Geographic and socio-demographic disparities in oncology care pathways

Eric Daoud

université PARIS-SACLAY

Geographic and socio-demographic disparities in oncology care pathways

Etude des disparités géographiques et socio-démographiques dans les parcours de soins en oncologie

Thèse de doctorat de l'université Paris-Saclay

École doctorale nŕ 582, Cancérologie : Biologie, Médecine et Santé (CBMS)
Spécialité de doctorat: Recherche clinique, innovation technologique, santé publique
Life Sciences and Health. Référent : Faculté de médecine

Thèse préparée dans les unités de recherche de l'**Institut Curie, équipe INSERM U932** et du **DI ENS - Département d'informatique de l'École Normale Supérieure, équipe DYOGENE** , sous la direction de **Fabien REYAL**, Professeur des Universités, et la co-direction de **Marc LELARGE**, Directeur de Recherche.
Thèse soutenue à l'INRIA Paris, le 12 Décembre 2022, par

Eric DAOUD-ATTOYAN

Composition du jury

Membres du jury avec voix délibérative

Gaël Varoquaux Directeur de Recherche, Université Paris-Saclay	Président du Jury
Gwenn Menvielle Directrice de Recherche, Inserm, Sorbonne Université, Paris, France	Rapporteur & Examinatrice
Pr. Raphaël Porcher Professeur des Universités, Université Paris-Cité	Rapporteur & Examinateur
Pr. François Alla Professeur des Universités, Université de Bordeaux	Examinateur
Pr. Stéphanie Allassonnière Professeur des Universités, Univeristé Paris-Cité	Examinatrice
Inès Vaz-Duarte-Luis Maitre de Conférence des Universités, Université Paris-Saclay	Examinatrice

THESE DE DOCTORAT
NNT : 2022UPASL085

Acknowledgements

Je remercie tout d'abord les membres du jury pour avoir accepté de suivre mon travail: Pr. François Alla, Pr. Stéphanie Allassonnière, Gaël Varoquaux et Inès Vaz-Duarte-Luis. Un grand merci également à mes deux rapporteurs, Pr. Raphaël Porcher et Gwenn Menvielle pour avoir pris le temps de relire mon manuscrit. Je suis honoré d'être encadré par ces grands médecins et chercheurs que je n'aurais jamais imaginé rencontrer.

Ensuite, un immense merci à mes deux superviseurs de thèse, Pr. Fabien Reyal et Marc Lelarge, pour m'avoir fait confiance dès le début et pour votre encadrement de qualité pendant ces trois ans. J'ai beaucoup appris grâce à vous et j'en ressors grandi. Merci Fabien pour tes innombrables intuitions, ta franchise et ton exigence, particulièrement sur la visualisation de données. Tu as su faire le pont entre la médecine et la recherche en informatique, qui m'a permis de garder en tête le besoin des patients tout au long de mes projets. Merci Marc pour ton calme, ta rigueur, ta pédagogie et ton immense expertise en Machine Learning notamment. Tu as accepté de me superviser en cours de thèse et je t'en suis très reconnaissant, ma thèse n'aurait pas été la même sans ton aide.

J'ai eu la chance de travailler dans deux équipes de recherche, le RT2 Lab de l'Institut Curie et l'équipe Dyogene de l'INRIA Paris. Mes premiers remerciements vont au RT2Lab. Tout d'abord merci à Elise et Beatriz, mes co-thésardes avec qui nous avons résolu bien des problèmes autant scientifiques qu'administratifs, et qui m'ont permis de garder le sourire pendant toute ma thèse. Ensuite, Judith sans qui je n'aurais jamais commencé cette thèse et qui m'a été d'excellent conseil tout au long de mon parcours. Merci à Anne-Sophie, pour son accompagnement, ses conseils et son expertise dans la recherche médicale. Merci à

Nadir, Lidia, Amyn, Dorian, Jeanne, Aullène, Marie, Zoé, Emile, Jade, Eva, Sara et Imane pour avoir apporté une nouvelle dynamique à l'équipe. Enfin merci à Floriane, Paul nouveaux thésards de l'équipe, pour leurs conseils et leur aide, je vous souhaite le meilleur. Un grand merci également à Dominique, sans qui je n'aurais jamais pu survivre administrativement à l'Institut Curie. Ensuite, mes remerciements vont à l'équipe Dyogene. Je remercie d'abord mes collègues de bureau, dit "le bureau des boloss". Luca, tu es la première personne que j'ai croisé à l'INRIA et je te remercie infiniment pour ton accueil, pour m'avoir intégré dans l'équipe, mais aussi pour nos footings et sessions musique. Matthieu, Mathieu et Bastien, je vous remercie pour votre bonne humeur quotidienne et votre générosité qui m'ont fait passer d'excellents moments dans ce bureau. Merci également à mes collègues Ilia, Thomas, Jakob, Lucas, David, Cedric, Antoine, Romain, Laurent et Kevin, brillants chercheurs d'une grande humilité. Enfin merci à Hélène, qui m'a aidé sur tous les sujets administratifs avec une gentillesse et une efficacité remarquable. Merci à Capgemini et à sa division Invent pour nous avoir accueilli dans leurs bureaux, et nous avoir permis de travaillé avec des équipes de qualité et de bénéficier de leur aide et de leur expertise. Merci à Charlotte, Olivier, Johan, Marc-Felix, Charles, Hakim, Louis, Karim, Émilie, François et Cléa. Merci aux chercheurs de l'Institut National du Cancer Philippe-Jean Bousquet, Christine Le Bihan et Sophie Houzard, pour m'avoir accompagné pendant ces trois ans, ainsi que pour tous vos précieux conseils. Merci aux membres du CBIO de m'avoir accueilli pendant le début de ma thèse, en particulier Chloé-Agathe Azencott, pour m'avoir guidé tout au long de ce travail.

Merci à tous les acteurs de la science ouverte notamment arXiv, medrXiv et SciHub pour faciliter l'accès à la recherche à travers le monde. Merci à la Boulangerie Moderne, pour ses madeleines en chocolat de qualité inégalée, que je recommande à tout Paris.

Merci à amis pour avoir toujours été là pour moi, et m'avoir encouragé tout au long de mon parcours et de ma vie: ceux rencontrés à l'ECAM Lyon, particulièrement Arnaud, Hubert, Jimmy et Lancelot; à Centrale, Anne-Laure et Tom; et ceux de ManoMano, dont Vincent, Sébastien, Antoine, Florent, et Maxime.

Un immense merci à ma famille, qui compte énormément pour moi et que j'espère ren-

dre fier. En particulier, merci à mes parents pour leur amour et leur éducation que j'ai parfois trouvé sévère mais qui s'est révélée juste. Merci à mon frère qui a toujours été à mes côtés. Merci à Maké, ma marraine et seconde mère, pour s'être occupée de moi comme un fils. Merci à Bernard, pour m'avoir conseillé et guidé, sans qui je n'aurais jamais fait d'études d'ingénieur. Merci à ma grand mère, pour sa tendresse et son amour.

Mes derniers remerciements vont à mon oncle Dikran et mon grand père René, qui ne m'auront malheureusement pas vu finir cette thèse. Dikran, j'admirais ta détermination et ta gentillesse. Papy, tu es pour moi le plus bel exemple de réussite et d'humanité. Vous me manquez tous les jours.

Contents

List of abbreviations

WHO World Health Organization

INSEE Institut national de la statistique et des etudes economiques

PCA Principal Component Analysis

SAE Statistiques Annuelles des Etablissements

CH Centre Hospitalier

CHR/U Centre Hospitalier Regional / Universitaire

CLCC Centre de Lutte Contre le Cancer

PSPH Participant au Service Public Hospitalier

EBNL Etablissement a But Non Lucratif

MCO Medecine, Chirurgie, Obstetrique

FCA Floating Catchment Area

2SFCA Two Step Floating Catchment Area

E2SFCA Enhanced Two Step Floating Catchment Area

LP Linear Programming

OT Optimal Transport

LSCP Location Set Covering Problem

MCLP Maximum Covering Location Problem

PSO Particle Swarm Optimization

CAMION Catchment Area MaximizatION

INCA Institut National du Cancer

IPCC Intergovernmental Panel on Climate Change

SiMCa Sinkhorn Matrix factorization with Capacity constraints

SA Spatial Accessibility

POI Point Of Interest

LAP Linear Assignment Problem

GHT Groupement Hospitalier de Territoire

PMSI Programme de Medicalisation des Systemes d'Information

SNDS Système National des Données de Santé

ATIH Agence technique de l'information sur l'hospitalisation

CO_2 Carbon dioxide

GHG Greenhouse Gas

ICT Information and Communication Technologies

GP General Practitioner

HRS Health Recommender System

GCN Graph Convolution Network

VGAE Variational Graph Auto-Encoder

French summary

Le cancer est l'une des principales causes de mortalité dans le monde, représentant près de 10 millions de décès en 2020. Selon l'Organisation mondiale de la santé, une personne sur cinq dans le monde développe un cancer au cours de sa vie. Les développements importants des traitements oncologiques observés ces dernières années ont amélioré les résultats pour les patients atteints de cancer. Bien que ces progrès aient un impact positif, ils ont augmenté la complexité de la prestation des soins. Pour faire face aux défis posés par cette complexité, des parcours de soins ont été introduits. Dans la littérature, les parcours de soins ont été définis comme ń une intervention complexe pour la prise de décision mutuelle et l'organisation des processus de soins pour un groupe bien défini de patients pendant une période bien définie ż. Un parcours de soins vise à renforcer la qualité des soins en améliorant les résultats des patients, en augmentant leur satisfaction et en optimisant l'utilisation des ressources. La littérature fait état de multiples preuves de disparités dans les parcours de santé et de soins, dont certaines sont dues à des facteurs externes tels que le statut socio-économique ou le lieu de résidence. Par exemple, le statut socioéconomique, reflété par le revenu, l'éducation ou la profession, exacerbe les problèmes de santé, y compris le cancer.

En France, l'Institut national du cancer (INCA) est l'agence d'État pour l'expertise sanitaire et scientifique en cancérologie, chargée de coordonner les actions de lutte contre le cancer. Depuis 2003, l'INCA produit des rapports contenant des recommandations nationales et des mesures visant à mobiliser les acteurs de santé publique autour de la prévention, du dépistage, de l'organisation des soins, de la recherche, du soutien aux patients et à leurs

familles, et de l'après-cancer. À ce jour, trois plans cancer ont été publiés et le dernier couvrait la période 2014-2019. Ce plan est largement focalisé sur les inégalités de prise en charge en oncologie, avec pour objectifs d'accroître les connaissances sur cette question et de lutter contre ce problème par des interventions concrètes.

Les données de vie réelle des patients représentent un volume d'informations sans précédent, actuellement sous-exploité. En particulier, en France, la sécurité sociale génère une grande base de données structurée à des fins administratives : le Système National des Données de Santé (SNDS). Le SNDS rassemble des données administratives complètes et actualisées sur 98% de la population française. L'exploitation du SNDS, à des fins de recherche, est une opportunité exceptionnelle d'élargir le champ de la recherche à l'amélioration des parcours de soins.

L'objectif de ce travail est d'exploiter les données de vie réelles des patients pour fournir des mesures et des outils permettant de lutter contre les disparités dans les parcours de soins en oncologie, en France. Nous avons choisi d'aborder en les disparités géographiques et socio-démographiques, et nous ne nous sommes pas concentrés sur un site de cancer spécifique. La principale source de données utilisée a été la base de données du PMSI, pour accéder aux données des hôpitaux et étudier les parcours de soins des patients. Nous avons limité l'analyse à l'année 2018, et n'avons pas étudié l'impact de la pandémie de COVID dans les parcours de soins. Chaque métrique et outil que nous avons développé au cours de cette thèse pourra être réutilisé dans d'autres travaux de recherche. Nous avons tout d'abord proposé une caractérisation de chaque centre de soins en France en termes de spécialisation oncologique. Ce label oncologique aidera les médecins, les patients, les chercheurs ou les professionnels de la santé publique à mieux évaluer les hôpitaux et leur répartition spatiale dans le pays. Deuxièmement, nous avons calculé un score d'accessibilité à l'oncologie, pour identifier les zones où les hôpitaux spécialisés en oncologie sont rares. Troisièmement, nous avons proposé un algorithme d'optimisation pour cibler les hôpitaux qui devraient être développés en priorité pour améliorer cette accessibilité. Quatrièmement, nous avons étudié les déplacements des patients entre leur commune de résidence et les hôpitaux qu'ils

visitent. Nous avons développé un indice de la charge de déplacement pour mesurer non seulement le déplacement en tant que distance, mais aussi en tant que combinaison de la distance, de la durée et de la sinuosité de la route. Nous avons également estimé l'empreinte carbone des déplacements de ces patients et simulé un scénario dans lequel chaque patient se rendrait au centre spécialisé le plus proche. Nous pensons qu'une plus grande transparence dans les soins oncologiques pourrait bénéficier aux patients et les aider, ainsi que leur médecin, à trouver l'hôpital le plus adapté situé à une distance raisonnable. Ainsi, nous avons construit une application web qui répertorie toutes les caractéristiques des hôpitaux, à la fois à destination des patients et des médecins. Enfin, nous avons développé un algorithme d'allocation basé sur le transport optimal pour diriger les patients vers un hôpital proche et adapté. Cependant, nous n'avons testé ce modèle que sur des données synthétiques, des recherches supplémentaires sont nécessaires pour l'appliquer à des données réelles de parcours de soins.

Chapter 1

Introduction

1.1 Preamble

1.1.1 Cancer epidemiology

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 [1]. According to the World Heath Organization, one in five people worldwide develop cancer during their lifetime. The GLOBOCAN 2020 database, produced by the International Agency for Research on Cancer gives estimates of incidence and mortality for 36 cancers, worldwide [2]. According to their statistics, the most common new cases of cancer in 2020 were: breast cancer, with 2.26 million cases; lung cancer, with 2.21 million cases; colon and rectum cancers, with 1.93 million cases; prostate cancer, with 1.41 million cases; skin cancer (non-melanoma), with 1.20 million cases; and stomach cancer, with 1.09 million cases. The most common causes of cancer death in 2020 were: lung cancer, with 1.80 million deaths; colon and rectum cancers, with 916,000 deaths; liver cancer, with 830,000 deaths; stomach cancer, with 769,000 deaths; and breast cancer with 685,000 deaths (Figure 1.2). Finally, each year, approximately 400,000 children develop cancer. The most common cancers vary between countries, but cervical cancer is the most common in 23 countries. The 7 most common cancers accounted for more than half of all the newly diagnosed cancer cases in

2020, as illustrated on Figure 1.1. For women, breast cancer is the most commonly diagnosed cancer and the leading cause of cancer death, followed by colorectal and lung cancer for incidence, and vice versa for mortality [2]. For men, Lung cancer is the most frequently occurring cancer and the leading cause of cancer death, followed by prostate and colorectal cancer for incidence and liver and colorectal cancer for mortality [2]. Cancer incidence rate and mortality rate are higher in countries with high Human Development Index (HDI) [2].

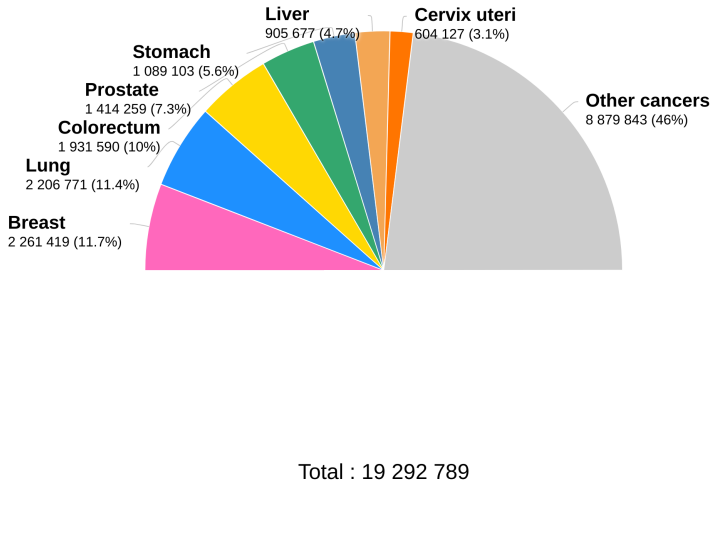


Figure 1.1: Estimated number of new cancer cases in 2020, worldwide, both sexes, all ages. According to the GLOBOCAN database, the most common cancers in terms of new cases were Breast, Lung, Colorectum, Prostate, Stomach, Liver and Cervix uteri. In 2020, these 7 cancer types represented more than half of all the newly diagnosed cancers.

1.1.2 What causes cancer

Official and trusted sources like the US National Cancer Institute or the World Health Organization have online resources explaining what is Cancer and what are its main causes. In the Cancer disease, some of the body's cells grow uncontrollably and spread to other parts of the body. Normally, the body's cells grow and multiply when the body needs them. Old

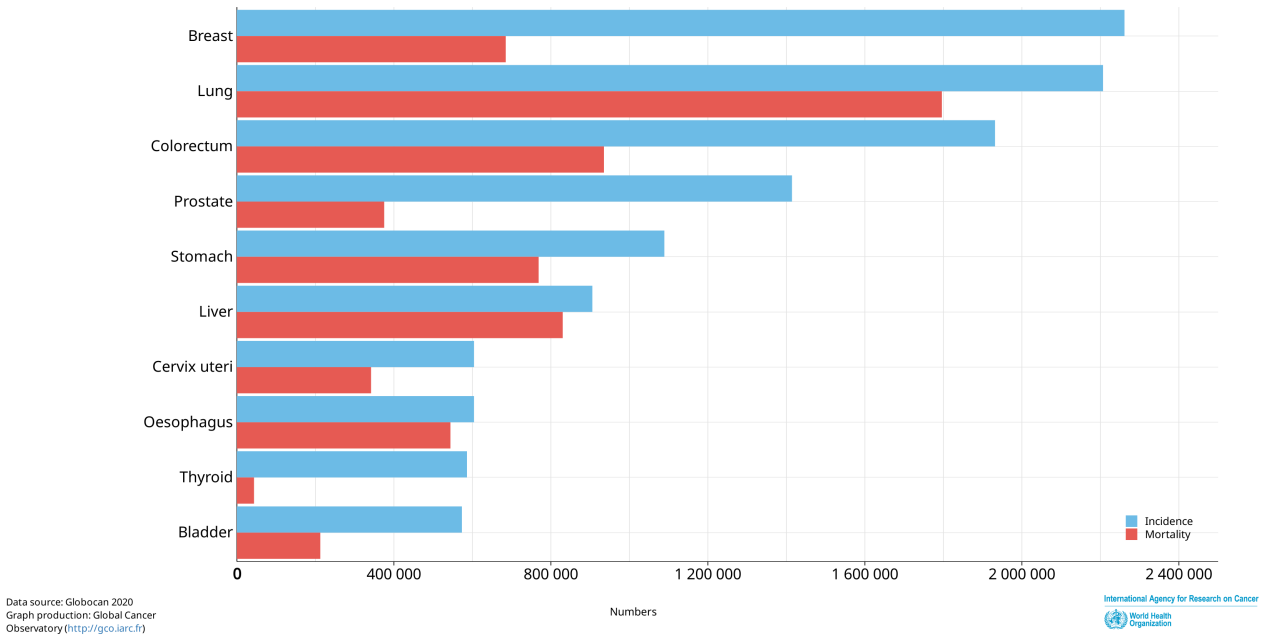


Figure 1.2: Estimated number of incident cases and deaths worldwide, both sexes, all ages. According to the GLOBOCAN database, the most common newly diagnosed cancers were Breat, Lung, Colorectum, Prostate, Stomach, Liver, Cervix Uteri, Oesophagus, Thyroid and Bladder. The most lethal cancers were Lung, Colorectum, Liver, Stomach, Breast, Oesophagus, Prostate, Cervix uteri, Bladder and Thyroid.

and damaged cells die and are replaced by new cells. This unwanted cells multiplication may form tumors, which are lumps of tissues. These tumors can either be cancerous (malignant) or non cancerous (benign). Malignant tumors can spread into nearby tissues or travel to distant places in the body to form new tumors, causing metatastic cancer. Cancer is a genetic disease, caused by changes to genes that control how our cells work. These changes can be caused by harmful substances in the environment, like chemicals in tobacco smoke or ultraviolet rays from the sun. They can also come from infections, or be inherited from our parents. The incidence of cancer rises dramatically with age. The overall risk accumulation is combined with the tendency for cellular repair mechanisms to be less effective as a person grows older.

1.1.3 Risk factors

A risk factor is defined as anything that increases the chance of developing a disease. While it is not possible to know when one will develop cancer, research shown that certain risk factors do increase the chances of developing cancer. Some risk factors include expose to chemicals, or certain behaviors like smoking. Some risk factors cannot be controlled, like age and family history. The most studied or suspected risk factors for cancer are: age; alcohol; cancer-causing substances; chronic inflammation; diet; hormones; immunosuppression; infectious agents; obesity; radiation sunlight; and tobacco.

1.1.4 Reducing the cancer burden

Research estimated that between 30% and 50% of cancers can be prevented by avoiding risk factors and implementing existing evidence-based prevention strategies. The following recommendations apply to minimize the cancer risk factors: not using tobacco; maintaining a healthy body weight; eating a healthy diet, including fruit and vegetables; doing physical activity on a regular basis; avoiding or reducing consumption of alcohol; getting vaccinated against HPV and hepatitis B if applicable; avoiding ultraviolet radiation exposure and/or using sun protection measures; ensuring safe and appropriate use of radiation in health care (for diagnostic and therapeutic purposes); minimizing occupational exposure to ionizing radiation; and reducing exposure to outdoor air pollution and indoor air pollution, including radon. Moreover, an early detection of cancer and the appropriate treatment and care can also reduce the cancer burden. As a matter of fact, for many cancers, the probability of being cured is high with an early diagnosis and the appropriate treatment. When identified early, the response to treatment is higher, as well as the survival probability. The treatments are usually less expensive. Significant improvements can be made in the lives of cancer patients by detecting cancer early and avoiding delays in care.

1.1.5 Cancer treatments

Every cancer type requires a specific treatment. A proper selection of treatment depends on both the cancer and the individual being treated. In most cases, the cancer treatment includes surgery; radiotherapy; and/or systemic therapy such as chemotherapy, hormonal treatments or targeted biological therapies. The primary goal of the treatment is either to cure cancer on considerably prolong life. Besides, maintaining a good quality of life is also important, and can be achieved with psychosocial and spiritual well-being or palliative care in terminal stages of cancer. We now explain briefly the the most common treatments for cancer. A comprehensive list with additional informations is available on the National Cancer Institute website.

- **Biomarker testing.** Biomarker testing aims at providing information on the individual's cancer, by looking for genes, proteins or other substances called biomarkers. As some biomarkers affect how cancer treatments work, biomarker testing is a way to choose the most suited treatment. Biomarker testing is an important part of precision medicine, in which the diagnosis and treatment are tailored to the to the genes, proteins, and other substances in the patient's body.

- **Chemotherapy.** Chemotherapy is a treatment that uses drugs to kill cancer cells. It aims at stopping or slowing down the growth of cancer cells. Chemotherapy is used to cure cancer, or ease the symptoms. While chemotherapy could be the only treatment received by patients, it is often administrated with other cancer treatments, based on the cancer type, the spread, and the other health problems (called co-morbidities). Chemotherapy treatment often introduces side effects such as mouth sores, nausea, hair loss and fatigue, the most common side effect. The induced fatigue is such that patients should be driven to and from chemotherapy; plan some rest on the day and the day after receiving it; and receive help for meals and childcare. Chemotherapy can be received during a hospital stay, at home, or an outpatient stay (no overnight). The treatment is administered in cycles: a period of chemotherapy treatment followed by

a period of rest.

- **Hormone therapy.** Hormone therapy slows or stops the growth of cancer that use hormones to grow, such as some prostate or breast cancers. Similarly to chemotherapy, hormone therapy is used to treat cancer or reduce its symptoms. Since hormone therapy interferes with hormones production, side effects may happen and can be different between men and women. Hormone therapy can be taken at home, in a doctor's office or in a hospital.

- **Immunotherapy.** Immunotherapy is a treatment that that helps the immune system fight cancer. As part of its normal function, the immune system detects and destroys abnormal cells and most likely prevents or curbs the growth of many cancers. For instance, immune cells are sometimes found in and around tumors. These cells, called tumor-infiltrating lymphocytes or TILs, are a sign that the immune system is responding to the tumor. People whose tumors contain TILs often do better than people whose tumors don't contain them. Even though the immune system can prevent or slow cancer growth, cancer cells have ways to avoid destruction by the immune system. Immunotherapy helps the immune system to better act against cancer. Several types of immunotherapy are used to treat cancer, including: Immune checkpoint inhibitors, T-cell transfer therapy, Monoclonal antibodies, Treatment vaccines, Immune system modulators. Immunotherapy drugs have been approved to treat many types of cancer. However, immunotherapy is not yet as widely used as surgery, chemotherapy, or radiation therapy. Immunotherapy can cause side effects, many of which happen when the immune system that has been revved-up to act against the cancer also acts against healthy cells and tissues in your body. You may receive immunotherapy in a doctor's office, clinic, or outpatient unit in a hospital.

- **Radiation therapy.** Radiation therapy, also called radiotherapy is a treatment that uses high doses of radiations to kill cancer cells and shrink tumors. These radiations damage the cells DNA, which will eventually stop dividing or die. The cells are not killed

right away. The treatment may last days or weeks before the cells are damaged. At that point, the cells will keep dying for weeks or months after the treatment ends. Most of the time, radiotherapy is given with other treatments such as surgery, chemotherapy and immunotherapy. Radiotherapy may affect nearby healthy cells, thus causing side effects.

- **Cancer surgery.** During this procedure, a surgeon removes the cancer from the patient body. Many types of cancer are treated with surgery, and it works best for solid tumors that are contained in one area. The surgery procedure can either remove the whole tumor, or part of it. It can also be used to ease symptoms, by removing tumors that are causing pain or pressure. The most frequent problems that can happen after surgery are pain and infection.

1.2 Subject definition

1.2.1 Care pathways

The important developments in oncology treatments seen in the recent years have improved outcomes for cancer patients. Even though these advances have a positive impact, they increased the complexity in the delivery of care. To face the challenges brought by this complexity, care pathways have been introduced. In the literature, care pathways have been defined as "a complex intervention for the mutual decision-making and organisation of care processes for a well-defined group of patients during a well-defined period" [3]. A care pathway aims at enhancing the quality of care by improving patient outcomes, increase patient satisfaction and optimizing the use of resources. Even though the adoption of care pathways is relatively new for some health services, the concept has long been existing, with first evidences during the 1950s [4]. In the recent years, care pathways have gained momentum, with multiple examples of adoption. Advantages of the care pathways include: faster diagnosis; greater consistency of care between providers, and better overview for patients; re-

ducing the risk of errors; and reduction of costs [4]. The expansion of treatment possibilities can lead to unwarranted variations that could affect the patients outcome. Adopting pathways is a way to ensure that all patients receive a consistent treatment, no matter where they are treated. Moreover, due to the sophistication of oncology care, most patients are treated by multi-disciplinary team of care providers, sometimes across different hospital sites. Care coordination is needed between all these providers, to avoid care gaps and potential errors. Again, pathways can facilitate the coordination by setting referral points, support data sharing and bring visibility to into treatment decisions made by all the care providers. Pathways could also benefit patients before they start their treatment, by promoting the appropriate use of precision oncology. For instance, by making sure that patients are not over or under-tested, or that the most optimal targeted therapeutic is selected based on the patient condition. Every process that aims at optimizing an operation should be monitored to identify and address under-performing areas. This is applies to care pathways as well. Storing and analyzing data related to treated patients during their patients would allow healthcare teams to evaluate their operations and optimize their practice patterns. Unwanted care variation could be quickly discovered and addressed, ultimately preserving patients.

However, due to the growing use of oncology pathways, some challenges have arisen, notably in the United States. The concerns included the process being used for pathway development, the administrative burdens on oncology practices of reporting on pathway adherence, and how to evaluate the true impact of pathway use on patient health outcomes [5]. As a result, the American Association of Clinical Oncology (ASCO) articulated a set of recommendations to improve the development of oncology pathways and processes. A total of 9 recommendations were proposed for clinical pathway development and implementation in the oncology setting. First, a collaborative and national approach should be pursued to reduce the administrative burdens associated with the non-managed proliferation of oncology care pathways. Second, the process for oncology pathways development should be consistent and transparent to all stakeholders. Third, the pathways should address the full spectrum of cancer care. Fourth, The pathways should be updated continuously based on

scientific knowledge, clinical experience and patient outcomes. Fifth, physicians should be allowed to easily diverge from pathways when evidence and patient needs dictate. Sixth, oncology pathways should be implemented in ways that promote administrative efficiencies for oncology providers and payers. Seventh, education, research and access to clinical trials should be promoted to patients during the pathways. Eighth, robust criteria should be developed to support certifications of oncology pathways. Lastly, research to understand the impact of pathways on patient outcomes should be supported.

1.2.2 Disparities in care pathways

There are multiple evidences of disparities in health and care pathways in the literature, with some due to external factors such as socioeconomic status or residence location. Socioeconomic status, reflected by income, education or occupation exacerbates health problems, including cancer [6]. An increase in mortality has been associated with lower socio-economic status. The cure rates of children with cancer are much higher in high-income countries than in the low-income ones [7]. Indeed, over 85% of children with cancer in high-income countries are cured, where only 20% in many low-income countries survive the disease. These disparities are caused by inadequate skilled workforce and health infrastructure. In colorectal cancer, people from low socioeconomic backgrounds had a higher incidence and mortality compared to other populations [8]. These disparities might result from differences in exposure to risk factors and limited access to prevention and treatment resources. In breast cancer, patients with low socioeconomic status experienced poorer survival rates after diagnosis due to more advanced cancer stage on presentation and poorer health condition [9]. Research also suggests that outcome disparities in breast cancer are due to differences in the quality of screening, diagnosis and treatment [10]. Overall, the outcome for all cancer sites combined was higher in poorer countries compared with more affluent countries. Poorer populations had 13% higher death rates in men, and 3% in women [11]. The rate difference between high and low socioeconomic populations urges the need for research into the mechanisms causing these disparities. Priority should be given to interventions

designed to reduce disparities by focusing on deprived populations since this is where the absolute differences in survival are [12]. A comprehensive literature review provided a list of disparities in cardio-oncology [13]. They classified these disparities into 4 social determinants categories: race and ethnicity; healthcare access a and quality; neighborhood and rurality; and economic stability. First race and ethnicity were shown to have an influence on outcomes, similarly to what has been discussed earlier. Second, poor healthcare access is linked to delayed care and worse outcomes. Then rurality was associated to worse outcomes compared to patients in metropolitan areas. Finally, poor economic stability results in a higher chance of renouncing to medical care. The research also suggests interventions to address these disparities, such as targeted policy intervention; increase diversity in clinical trials; increase access to cardio-oncology care; better resource allocation; use of social media to promote health literacy; and the integration of social determinants of health in clinical care delivery. Despite all these evidences, it seems that the allocation of healthcare resources is still mostly going to treat diseases, rather than addressing the predisposing factors of these inequalities [14].

Finally, gender appears to have an impact on care pathways. For example, men may have difficulty talking about their symptoms, fearing that it will be perceived as a sign of weakness; whereas women who require care are more likely to be neglected [15]. Women with myocardial infarction have a higher mortality rate than men, and this discrepancy appears to be partially due to delayed diagnosis and access to appropriate care [16]. Similarly, a pediatric study of kidney transplantation showed that young girls had less rapid access to transplantation than young boys. This is partly due to non-medical reasons such as parental and practitioner behavior regarding organ donation [17]. For Head and Neck Cancer (HNC), research found that women had an increased relative hazard ratio for death versus other causes compared with men. However, they were less likely to receive intensive chemotherapy and radiotherapy than men. This might indicate that women in this cohort may be under-treated in clinical practice and potentially miss the opportunity for their HNC to be aggressively treated [18]. Lastly, women have been under-represented in clinical trials. Al-

though enrollment of women has increased over time, it remains lower than the relative proportion in the disease population [19]. Overall, the gender of the patient could have an impact on the oncology care pathway. Indeed, several studies show that women's treatment for several types of cancers is suboptimal. This would at least partially explain why their chances of survival from these diseases are lower than those of men [18, 20, 21]. The above examples suggest that patient survival could be improved by taking gender into consideration in the care pathway. However, at present, gender differences in the oncology care pathway are barely explored.

1.2.3 Cancer in France

In France, The National Cancer Institute - Institut National du Cancer (INCA) is the State agency for health and scientific expertise in cancer, responsible for coordinating actions to fight cancer. Created by the Public Health Law of August 9, 2004, it is placed under the joint supervision of the Ministry of Solidarity and Health on the one hand, and the Ministry of Higher Education, Research and Innovation on the other. Since 2003, INCA produces reports with national recommendations and measures to mobilize public health actors around prevention, screening, organization of care, research, support for patients and their families, and post-cancer care. These reports are called "cancer plans", and are supported at the highest government level in the country by the President. To date, three cancer plans have been issued, and the last one covered the 2014-2019 period [22]. This plan is largely focused on inequalities in oncology care, with objectives to increase knowledge on this issue and address the problem through specified interventions. These inequalities in cancer care cover multiple dimensions. Some are territorial; others are social and environmental; and also depend on other factors such as the age of individuals, their sex or their genetic characteristics. Inequalities also exist in access to and use of screening, treatment and therapeutic innovation. This is reflected in particular in the fact that diagnoses are often made later for disadvantaged social groups, leading to lower outcomes and more invasive treatments. Similarly, people with lower incomes or living in deprived communities experience longer

delays in entering the healthcare system, or between the different phases of this system. Understanding where the inequalities are coming from is a required step to propose working solutions. The report mention that an information system to monitor health inequalities was lacking and should be developed. Matching socio-demographic databases with cancer observation and surveillance tools is endorsed by the cancer plan. Moreover, the regional health agencies should be regularly provided with territorialized data on cancer inequalities. Overall, this last cancer plan provides support for research in cancer inequalities and population health. The fight against inequalities in cancer care goes far beyond the field of health. This issue must mobilize actors from the social sector as well as from education and research. All levels of public action are concerned, from local to national. This public policy must be built by systematically integrating the point of view and expertise of patients, especially those from the least privileged social categories. We should develop solutions to improve their involvement in the processes and approaches of health democracy.

1.2.4 Leveraging medical data in France

SNDS database

Real-life patient data represents an unprecedented and currently underutilized volume of information, currently under-exploited. In particular, in France, the Social Security generates a large structured database for administrative purposes: the National Health Data System - Système National des Données de Santé (SNDS). The SNDS gathers complete and up-to-date administrative data on 98% of the French population. The exploitation of the SNDS, for research purposes is an exceptional opportunity to broaden the scope of research in the improvement of care pathways. Indeed, the substantial number of patients it contains exceeds the size of all the French cohorts collected in the treatment centers. The SNDS is one of the largest health databases in the world. It attracts research thanks to its almost complete coverage of the French population, which makes it possible to work on the complete care pathway of patients. A major challenge for the SNDS is to make these data available to promote studies, research or evaluations of public interest. The SNDS has been effectively

used for research on the following topics: information on health and health care supply; evaluation of health policies; evaluation of health care expenditure; information of health professionals on their activity; health monitoring and safety; research, studies, evaluation and innovation in health. The SNDS databases contain notably the following data sources: health insurance data; hospital admissions data; and medical causes of death; Overall, the SNDS contains more than 3,000 variables; an annual flow of 1.2 billion health records; 11 million hospital stays; 500 million procedures; and 450 TB of data. The SNDS contains notably the following data: expenditures and reimbursements; prescriptions (drugs); medical devices; usage of other services such as transport; hospital activity and stays; daily allowances and long-term conditions. The patients characteristics stored in the system are their age, sex and municipality residence. Patients can be followed throughout their pathways by a unique identifier. The data in the SNDS are kept for a total of 20 years, then archived for 10 years. However, the SNDS does not contain: clinical examination results such as imaging or biological data; paraclinical data such as smoking, blood pressure, BMI; medical consultation reasons; risk factors such as tobacco, alcohol, or nutrition; drugs delivered during hospital stays; social data.

PMSI database

The Programme de Medicalisation des Systemes d'Information (PMSI) is a database part of the SNDS, focused on hospital data. It provides a synthetic and standardized description of the medical activity of almost every hospital in France. The PMSI model was imported from Boston, MA, from Professor Robert Fetter (Yale University) and the DRG (Diagnosis Related Groups) models. It was an empirical construction of hospitalization costs based on several million hospital stays. Initially, in France, it was used only for descriptive purposes, and not for financial purposes. The PMSI was gradually extended with experiments in both the public and private sectors. The purpose of these experiments was to study the feasibility of pricing based on the PMSI. Since 2005, it has been used for the implementation of activity-based pricing (T2A), a new system for remunerating hospitals based on their activity. The PMSI

database is used within 4 pans of the hospital activity: medicine, surgery, and obstetrics (MCO); psychiatry, follow-up care; and home hospitalization. We restricted all our analyses to the MCO section.

The PMSI MCO database is populated with data gathered in the hospital. For each stay of an inpatient, a standardized discharge summary (RSS) is produced. This RSS is produced as soon as possible after the patient's discharge. It must contain a main diagnosis, which is the pathology that motivated the patient's admission to the medical unit (UM). The RSS can also contain a related diagnosis, describing the reason for the stay, and associated diagnoses. The related diagnosis role is to improve the documentary accuracy of the coding. Diagnoses are coded according to the ICD-10 (International Classification of Diseases and Use of Health Services No. 10) published by the World Health Organization (WHO) and regularly extended by the French Ministry of Health. It may also contain technical procedures coded according to the CCAM (Common Classification of Medical Procedures). Each care unit during the stay provides a medical unit summary (RUM) at the patient's discharge. The RUM contains data concerning the patient's stay in a given UM: patient's date of birth, gender, municipality, date of entry into the UM, date of discharge and his medical data such as the main diagnosis, associated diagnoses and procedures With the synthesis of the successive RUMs, the standardized discharge summary (RSS) is produced for the whole stay. The RSS are anonymized and then become anonymous discharge summaries (RSA) for transmission to the regional health agency (ARS).

1.3 Objectives and contributions of the thesis

1.3.1 Objectives

The objectives of this work is to leverage the real-life patient data to provide metrics and tools to address the disparities in oncology care pathways, in France. We chose to address the geographic and socio-demographic disparities first, and we did not focus on a specific cancer site. The principal data source used was the PMSI database, to access hospitals data

and study the patients care pathways. We restricted the analysis to the year 2018, and did not study the impact of COVID pandemic in the care pathways. Every metric and tool that we developed during this thesis will be available for reuse in other research works.

1.3.2 Organization of the thesis

The following paragraphs describe the chapters of this thesis. First, we proposed a characterization of every care center is France in terms of oncology specialization. This oncology label will help physicians, patients, researchers or public health professionals to better evaluate the hospitals and their spatial distribution in the country. Second, we computed an oncology accessibility score, to identify areas where oncology dedicated hospitals are scarce. Third, we proposed an optimization algorithm to target the hospitals that should be developed in priority to improve the oncology accessibility. Fourth, we studied the patients travel from their municipalities of residence to the hospitals they visit. We developed a travel burden index to not only measure travel as a distance, but as a combination of distance, duration, and road sinuosity. We also estimated the carbon footprint of these patients travel, and simulated a scenario where every patient would travel to the nearest specialized center. Fifth, we argued that more transparency in oncology care could benefit patients and help them and their physician to find the most suited hospital located in a reasonable distance. We built a web application that lists all the hospitals characteristics, for both patients and physicians. Finally, we developed an allocation algorithm based on Optimal Transport to address patients to a nearby and suited hospital. However we only tested this model on synthetic data, more research is needed to bring it to actual pathways data.

Care center characterization

Countries, such as the UK, USA and Canada, have been implementing a policy of centralizing the care of patients for many specialized services [23]. With such policy, patients are directed to a limited number of hospitals with higher volumes and more specialized surgeons. There is evidence that this process will have a positive impact on the health outcomes of those

patients treated in these specialized centres. For instance, centralized care is beneficial for patients undergoing high-risk procedures, these surgeries have lower mortality rates when performed by high-volume surgeons [24, 25, 26, 27, 28]. A centralized service for ovarian cancer may lead to better survival outcomes; evidence from various other sources suggests that this may also be more cost-effective [29]. With the rural exodus, the sparsely populated areas expanded, and several hospitals are serving relatively small populations. As a result, surgeons operating in these facilities are managing fewer cases of a given disease. For instance, in the South West of England, surgeons treating epithelial ovarian cancer were managing fewer than ten cases of ovarian cancer per year. There is a need to maintain a critical volume of work in order to sustain surgical expertise [30].

Through all these evidences, it is clear that not all the hospitals are equal for cancer treatment. In France, there are many hospitals that do not have the same degree of oncology specialization. Hospitals are classified into different legal categories like public hospitals or private structures, but there is no indicator to assess the degree of oncology specialization and how large the hospital is. In this chapter, we first proposed a method to automatically label all the hospitals in metropolitan France, based on their statistics and available health services. Lastly, we studied the collaborations between the hospitals, based on patients who visited multiple hospitals during their pathways. Through community detection algorithms, we grouped hospitals that frequently exchange patients together. By adding the oncology specialization label within the discovered communities, we believe we can propose new hospital groups that are based on patient real-life data, to improve collaborations and ultimately benefit the patients. In this chapter, we first proposed a method to automatically label all the hospitals in metropolitan France, based on their statistics and available health services. Lastly, we studied the collaborations between the hospitals, based on patients who visited multiple hospitals during their pathways. Through community detection algorithms, we grouped hospitals that frequently exchange patients together. By adding the oncology specialization label within the discovered communities, we believe we can propose new hospital groups that are based on patient real-life data, to improve collaborations and ultimately

benefit the patients.

Accessibility score

While a lot of the ongoing research is focusing on finding new cancer treatments, accessibility to oncology care receives less attention. Accessibility refers to the relative ease by which services can be reached from a given location [31]. Accessibility can be defined by spatial factors, determined by where you are; and non-spatial factors, determined by who you are [32]. In what follows, we restrict accessibility to Spatial Accessibility (SA) and use both terms interchangeably. SA methods assess the availability of supply locations from demand locations, connected by a travel impedance metric. Supply locations are characterized by their capacity or quantity of available resource. Similarly, demand locations are characterized by their population. Such methods have been successfully used to measure access to healthcare, such as primary care [33] or oncology care [31, 34, 35] in several countries including France [36, 37, 38]. When measuring accessibility for healthcare, the supply locations are often physicians locations, whose capacity might be the number of physicians at that location. Population locations represent where patients live. This could be the precise address or a municipality. However, while accessibility to primary care have been described in several studies, there is little work that focused on oncology care specifically. In what follows, we applied SA methods to quantify the accessibility the oncology care in metropolitan France. Intuitively, we compute a score for every municipality that measures how easy it would be for patients living in a given municipality to reach oncology care.

In what follows, we applied SA methods to quantify the accessibility the oncology care in metropolitan France. Intuitively, we compute a score for every municipality that measures how easy it would be for patients living in a given municipality to reach oncology care.

Accessibility optimization

Uneven distributions of population and health-care providers lead to geographic disparity in accessibility for patients [39], illustrated by our previous results on accessibility. Several

methods have been developed to address these disparities. Location-allocation algorithms [40] can optimize the distribution and supply of health providers to reduce accessibility disparities. These algorithms seek the optimal placement of facilities for a desirable objective under certain constraints [31]. For instance, an optimization algorithm was developed to improve the healthcare planning in rural China by finding the best place and capacity for new health facilities [41]. A spatial optimization model was designed to maximize equity in accessibility to residential care facility in Beijing, China [42]. When optimizing health accessibility, there are two competing goals: equity and efficiency [43, 44]. Equity may be defined as equal access to healthcare for everyone [45]. An efficient situation is when everything has been done to help any person without harming anyone else [46]. While some argue that efficiency should be ad-dressed in priority [46], others agree that equity is a matter of ethical obligation, especially in public health [47, 48]. Regarding efficiency optimization, the most popular algorithms are p-median, Location Set Covering Problem (LSCP) and Maximum Covering Location Problem (MCLP). The p-median algorithm minimizes the weighted sum of distances between users and facilities [49]. LSCP minimizes the number of facilities needed to cover all demand [50]. LSCP maximizes the demand covered within a desired distance or time threshold by locating a given number of facilities [51]. To reach equal access to healthcare, quadratic programming has been used to minimize the variance of accessibility scores defined by the Two Step Floating Catchment Area (2SFCA) [52]. Similarly, a Particle Swarm Optimization (PSO) algorithm was developed to minimize the total square difference between the accessibility score of each demand location and the weighted average accessibility score [42]. Finally, a two-step optimization algorithm has been developed to address the dual objectives of efficiency and equality, by first choosing where to site new hospitals and then deciding which capacity they should have [53, 54].

However, most of the previous algorithms seek locations to open new health facilities. Regarding oncology care, opening new facilities can be very costly and hard in practice. In this work, we are interested in the case where the health facilities locations are fixed, and the only lever to improve accessibility is to increase their capacities.Given a capacity bud-

get, we want to know which facilities to grow and by how much. We introduce CAMION, an accessibility optimization algorithm based on Floating Catchment Area (FCA) and Linear Programming (LP). The initial accessibility score was computed with the Enhanced Two Step Floating Catchment Area (E2SFCA) algorithm [55] but our algorithm can generalize to more FCA derivatives. In the following sections, we proposed two approaches for optimizing the accessibility scores. The first one is an overall optimization, where we seek to maximize the total accessibility. The second one is a maxi-min optimization, where we want to maximize the minimum accessibility instead. The first approach could be seen as efficiency maximization where the second method aims towards equity. Then, we embedded our results and algorithms into a web application called "oncology-accessibility". Through this web application, we let the users run the optimization algorithm with the parameters they want, and visualize the output on interactive maps and figures. We believe such an app could benefit the healthcare professionals, to help addressing the accessibility disparities in the country.

Patients routes

Cancer treatment delay is a problem in health systems worldwide, increasing mortality for many types of cancers [56], including breast cancer [57, 58, 59]. Distance between patients residence and diagnosing hospitals is among the factors causing these delays, especially for cancer types that are hard to diagnose [60]. While accessibility to healthcare is growing, research found that 8.9% of the global population (646 million people) could not reach healthcare within one hour if they had access to motorized transport [61]. Thus, a non insignificant part of the population might be exposed to lower prognosis.

The benefits of centralized healthcare have been debated. A centralized approach often requires patients to travel far away from their home and their local community hospitals [29]. Patients subject to longer travels to reach a specialized hospital are likely to be affected by the travel burden and separation from their social environment [62]. In the debate between local versus centralized healthcare provision, there are evidence of an association between travel distance and health outcomes [23]. Unsurprisingly, travel to cancer treat-

ment is inconvenient for some patients and might even act as a barrier to treatment [62]. Research also showed that patients who lived far from hospitals and had to travel more than 50 miles had a more advanced stage at diagnosis, lower adherence to encoded treatments, a worse prognosis, and a worse quality of life [63]. More research linked travel burden with lower treatment compliance [64, 65]. The distance from the hospital influences the choice of appropriate treatment by cancer patients. In breast cancer, patients living farther from a radiation treatment facility more often underwent mastectomy instead of breast conservative surgery [66, 67, 68, 69, 70, 71] or did not undergo radiotherapy after breast cancer surgery [72, 66, 67]. In non small cell lung cancer, patients were most likely to not undergo potentially curative surgery if they lived far from a specialist hospital and only attended a general hospital for their care [73]. Moreover, the necessity for repeated visits for cancer diagnosis and treatment makes distance an even more important issue for the patient[65]. However, for hard to diagnose cancer type like rectum or testis cancers, distance was associated with decreasing odds of advanced disease stage [74]. This is possibly due to being treated in more specialized hospitals. The negative effects of centralized healthcare are even more pronounced for patients living in rural areas. Indeed, rural cancer patients face more challenges in receiving care, due to the limited availability of providers and clinical trials, as well as transportation barriers and financial issues [75]. There are evidence of poorer treatments and outcomes for patients living in rural areas. For instance, in Australia, poorer survival and variations in clinical management have been reported for breast cancer women living in non metropolitan areas [76]. Still in Australia, breast cancer women treated in a rural hospital had a reduced likelihood of breast conservative surgery [77]. The hazard of death from ovarian cancer was greater in women treated at a public general hospital than in women treated at a gynecological oncology service [78]. Contacting a provincial hospital instead of a university hospital might lead to diagnosis and treatment delays, which could be improved by a better referral system [79]. In Australia, patients living farther from a radiotherapy service were more likely to die of rectal cancer, with a 6% risk increase for each additional 100km [80]. In Rwanda, rural breast cancer patients who lived in the same dis-

trict as breast cancer hospitals had a decreased likelihood of system delay [59]. In Canada, place of residence seems to influence health outcomes in patients with diffuse large B-cell lymphoma [81]. They found that rural and metropolitan patients had similar survival; however, patients in small and medium urban areas experienced worse outcomes than those in metropolitan areas. Thus, rural culture might have a dual effect on health outcomes. On one hand, distance, transportation, and health services shortage are barriers to healthcare. On the other hand, rural culture comes with community belonging, and deeper relationship with health care professionals, which might be beneficial for some patients [82].

Additionally to having a negative impact on patients health, longer travels participate in global warming due to their Carbon dioxide (CO_2) emissions. The World Health Organization called climate change the greatest threat to global health in the 21st century, significantly affecting hundreds of millions of people [83]. The United Nations created the Intergovernmental Panel on Climate Change (IPCC) to assess the science related to climate change and provide governments with scientific information that they can use to develop climate policies. The health care sector is an important contributor to CO_2 emissions. An international comparison of health care carbon footprints showed that, on average, the health carbon footprint in 2014 constituted 5.5% of the total national carbon footprint [84]. Hence, the health sector has a responsibility to take climate action [85]. Especially since the Paris Agreement, where countries agreed to cut Greenhouse Gas (GHG) emissions to keep global warming below 2 degrees Celsius. Today, hospitals are powered by fossile energy such as coal, oil and gas. Healthcare related travels, and the manufacture and transport of healthcare products are also major causes of GHG emissions. Ultimately, all health systems will need to reach near zero emissions by 2050, which can be more cost effective than business as usual. The Lancet Countdown on health and climate change started to review annually the relation between health and climate change [86]. A large share of these carbon emissions is due to patients journeys [87, 88] because most patients travel by car [89]. With centralization of care, patients are encouraged to be treated in large hospitals for better outcome [90]. Such hospitals are in urban areas, and the populations living in rural areas will have to travel

longer to reach these centers, resulting in higher carbon emissions. In France, few studies have evaluated the ecological impact of cancer care [91]. The Shift Project is a French think tank that works towards a carbon-free economy. As a non-profit organization, they inform and influence the debate on the energy transition. In 2021, the Shift Project released a report on how to decarbonize the health care sector in France [92]. They identified that most of the GHG emissions were scope 3 emissions, which are indirect emissions that occur in the hospitals value chain. Among these emissions, the largest source are pharmaceuticals and medical device buying, followed by patients and visitors transportation. The Shift Project states that emissions related to transportation should be cut by 99%, through measures like increasing public transportation and telemedicine. Telemedicine includes all medical practices that allow patients to be treated remotely from a health facility. It has been used increasingly around the world, even in oncology where it is sometimes referred as teleoncology [93, 94, 95, 96]. Teleoncology models have been used to provide access to specialized cancer care for people in rural, remote and other disadvantaged areas, which minimizes the access difficulties and disparities [97, 94]. Teleoncology models can also be beneficial in training medical, nursing, and allied health trainees and staff at rural centers [96]. Research reported multiple benefits of telemedicine at every level of care, including education, prevention, diagnosis, treatment, and monitoring [98]. However, besides the expected benefits, several questions and fears are emerging [98]. First, there is a risk of patient isolation, due to the absence of in-person meeting. It is also more difficult to build an atmosphere of trust during remote consultations and the examinations might be of inferior quality. Finally, digital divide is a major limitation of e-health, as certain categories of patients do not have access to the internet or to a smartphone.

In this chapter, we analyzed the travels of cancer patients in metropolitan France. Our goal was to assess whether the earlier observations on the negative effects of centralization of care were happening in France. Hence, we first described the travel duration distribution in metropolitan France, and compared it with the population densities and the oncology specialization of the visited hospital. Then, we argued that the negative effects of travel on

cancer patients was not only due to driving distance and duration: the road sinuosity should also be taken into account. We proposed a travel burden index, which is a composite indicator based on multiple variables to evaluate how easy it is to go from a population location to an hospital. Additionally, we estimated the carbon footprint of cancer patients travels, and compared these numbers across the different regions. Finally, we ran an optimization algorithm to simulate the scenario where every patient traveled to the closest hospital, such that the hospitals capacities were not exceeded. We only considered Breast Cancer patients as this cancer is relatively frequent, and many hospitals have the required expertise.

Transparent healthcare

Over the past few years, there has been a massive change in the way we communicate and interact with information. The amount of data and content available to the public keeps increasing, as well as the number of information delivery platforms. Studies define this phenomenon as "the communications revolution" [99]. Smartphones democratization and adoption rate are partly responsible for this revolution. Indeed, a large and growing number of people own a smartphone, enabling them to access information anytime and anywhere. Through this, there has been a change in how people access and use information. With the increasing number of media sources, mass audience is now split into smaller groups who share common characteristics and interests. Also, the growth of online audience is now far outpacing the other media. As a benefit of this communication revolution, it is getting easier to access resources online, even technical resources such as technical reports and scientific articles. While these materials may not always be intended for a mainstream audience, their availability offers opportunities for access and interpretation by different groups.

The healthcare sector is no exception in this revolution, and health resources are increasingly available online [100, 99], changing how patients interact with health providers. Communication has been found to play a central role in cancer prevention and control. It can provide information on cancer prevention, monitor lifestyles and health behaviors, promote participatory decision making during cancer detection, diagnosis, and treatment, and

foster quality of life during survivorship or end of life [99]. When diagnosed with cancer patients and their family members lives change radically. They receive treatments and have to make choices with serious consequences. Such diseases and treatments are complex, but should be understood before decisions are made. Patients and their family members should be provided with intelligible and up to date information on the stage of disease, treatment options and complementary therapies [101, 102]. Multiple benefits of bringing more information to the patients have been reported. Involving cancer patients in decision-making on their pathways improves their satisfaction and quality of life, compliance with treatment and their ability to manage symptoms [103, 104, 105, 106, 107, 108, 109, 100]. Moreover, medically related education interventions are most effective when they are tailored to patients' individual needs, especially for cancer patients [109]. Through all these benefits, it is clear that monitoring patient information seeking experiences over time is important [110].

As a matter of fact, patients are often seeking information during their pathways. In the United States of America, a survey from the Health Information National Trends (HINTS) [111] measured online health activities, levels of trust, and source preference for 6,369 people. They observed that physicians remained the most trusted source of information, despite an increasing number of people looking for information online. However, there is increasing evidence in the literature that patients are often not satisfied with the information they received. Some reported to lack information on their disease and its consequences [106], while others forget or misunderstand the information conveyed [112, 113]. The interaction with their physician has also been cited as a major cause of dissatisfaction [114, 115] at all stages of illness [116]. Patients reported insufficient time spent on communication during the clinical encounter and physicians inability to keep up with the most current information and advances in cancer care [117]. Some patients reported incorrect diagnosis, or not receiving the most up-to-date cancer information from their physician, especially for rare cancers [118]. Patients who need health information but experience difficulties have been found at risk of experiencing poorer psychosocial health [119]. A Canadian study surveyed patients attending appointments at follow-up cancer clinics in Calgary, Alberta [107] between 2011

and 2012. They approached 648 patients and obtained responses from 411 one of them. The study aimed at: identifying information needs of patients when meeting their physician for a follow-up; listing patients preferences on how to receive information. Here are the results they gathered regarding information seeking patterns. The most frequently reported source of information was the Internet (57.4%); health provider (32.6%), brochures or pamphlets (25.1%), and cancer organizations (24.3%). The most frequently reported types of information sought included information about a specific type of cancer (43.1%), treatment or cures for cancer (29.4%), prognosis or recovery from cancer (29.0%), and prevention of cancer (27.0%). The least frequently reported types of cancer information sought included where to get medical care (3.4%), paying for medical care or insurance (4.6%), and cancer organizations (5.4%). Regarding trust, the physician or health care provider was largely the most trusted source of information, followed by Internet, and family and friends. The least trusted sources of information included radio, newspaper, and television. More evidence is reported on the use of the internet for health information retrieval [120, 121, 122, 118]. For instance, an online questionnaire was administered to participants of cancer-related communities hosted by the Association of Cancer Online Resources (ACOR) [118]. As a result, 488 participants shared their personal experiences on why and how they accessed online health resources. Participants who experienced a lack of informational support related to procedures found blogs and testimonies online that helped them to know what to expect from a physical and emotional perspective. Moreover, for rare diseases, physicians might actually benefit from patients looking from additional information online, as it could bring additional knowledge to them, and even change their plans for care. Aware patients can also challenge their physicians by asking meaningful questions and participate in the tailoring of their treatment plans. Finally, online communities allowed patients to identify physicians with a proven track record in cancer care. They endorsed care providers who took the time to answer questions, as well as specialists from major cancer centers, that brought superior care which led to better outcome. Indeed, General Practitioner (GP) play a crucial role in early cancer detection because the majority of cancer patients initially consult their GP with

symptoms. Therefore, the actions taken by the GP upon the patient's symptom presentation may considerably affect the cancer trajectory [60]. To sum up, the increased usage of the Internet by cancer patients puts new demands on health care professionals. Patients need advice about how to find reliable and credible web sites and also help with authenticating and interpreting the information they find [123].

While patients are looking for informations on their symptoms, diseases and treatments, it would be crucial for them to know better about their physician's ability, especially for cancer surgery. In cancer care, surgery is one of the most important part of the treatment, and is directly linked to the surgeon ability. Surgeon and hospital-related factors have been found to be direct predictors of outcome in colorectal cancer surgery [124, 125]. In breast cancer, patients managed by high-volume surgeons were more likely to have breast-conserving surgery (BCS) than those managed by low-volume surgeons [126]. Moreover, breast cancer patients who receive treatment from experienced and specialized surgeons are more likely to receive the standard sentinel lymph node biopsy [127]. The surgeon's expertise and learning curve is directly related to the patient's outcome [128]. A low surgeon or hospital caseload may be compensated by intensified supervision or by improved training and teaching [125]. From all these findings, it is questioned whether surgeons should have an ethical obligation to inform patients of their surgical volume and outcomes [129]. One way to monitor the surgeons abilities is the use of quality indicators, which have been developed in high income countries and contributed to improved quality of care and patient outcomes over time [130]. With these evidences of healthcare information needs, we developed Healthcare Network, a web application that lists every hospital in France, and displays key statistics on them. The application is directed to either health professionals or patients. Health professionals might use it to gain insights about specific hospitals, and look for the best place to send their patients when they lack expertise. Patients could learn more about the hospital they have been sent to, check the care quality or surgery volume.

Sinkhorn Matrix factorization with Capacity constraints

In a very broad range of applications – many of them being led by e-commerce leaders (Amazon [131], Netflix [132]) – recommendation algorithms have been increasingly used over the past decade. These algorithms are capable of showing users a personalized selection of items they may like, based on their interests and user behavior.

Up to now, the predictions are built upon user-item affinity scores (e.g., user/movie ratings) which are obtained from high-dimensional embeddings of items and users. While these approaches work for most e-commerce applications, there are other natural settings in which more attributes should be considered in the recommendation process. For instance, item capacity constraints are of paramount importance in location or route recommendation, where recommending the same item to every user could lead to congestion and significantly deteriorate user experience [133]. Moreover, in the case of location recommendation, travel distance is also a key factor: the user's choice is often the result of a trade-off between affinity and proximity [134]. In the healthcare sector, patients are usually addressed to an hospital by their general practitioner – or by word of mouth. Since the choice of hospital and practitioner may be critical, an important issue is to make sure that patients are routed to the best place possible – namely to a nearby and adapted structure, without capacity saturation. Benefits of such systems have been documented in the literature. For instance, an application similar to Google Maps for guiding patients to different care centers in a multi-site hospital reduced patient travel time [135]. Another research proposed a method to select the optimal care center using several criteria such as geographic accessibility and service quality. In particular, transportation networks such as high-speed lines and highways are taken into account in the center selection [136].

In this work, we study the recommendation problem in the setting where affinities between users and items are based both on their embeddings embeddings in a latent space and on their geographical distance in their underlying euclidean space (e.g., $\mathbb{R}^2$), together with item capacity constraints. Upon the observation of an optimal allocation, user embeddings, items capacities, and their positions in the euclidean space, our aim is to recover item

embeddings in the latent space; doing so, we are then able to use this estimate e.g. in order to predict future allocations. Our contributions are as follows:

(i) we propose an algorithm based on matrix factorization enhanced with optimal transport steps to model user-item affinities and learn item embeddings from observed data;

(ii) we then illustrate and discuss the results of such an approach for hospital recommendation on synthetic data.

After reviewing related work, we formally define the problem in mathematical terms, we describe our algorithm for SiMCa and give theoretical guarantees on its convergence. We then illustrate our method for the hospital recommendation problem on synthetic data, discussing the results as well as the choice of parameters.

Chapter 2

Care centers characterization

This chapter is part of a research article currently under submission. The-preprint is available on medrXiv.

2.1 Methods

2.1.1 Labelling hospitals by oncology specialization

Data collection

In this section, we detail how we gathered the data collection process to run our method. We first needed health data to characterize the care centers. Then, geographical and socio-demographic data was used to obtain information on the population locations. Health data is collected from two sources: the PMSI and the Statistiques Annuelles des Etablissements (SAE) databases. The PMSI database is includes discharge summaries for all inpatients admitted to public and private hospitals in France. The SAE database is a compulsory and exhaustive administrative survey of all public and private hospitals in France. The survey is sent every year and describes the activities of the hospitals as well as the list of services and activities they provide. The list of hospitals in France is available in the PMSI database and updated yearly. There were 5,148 hospitals in 2018. To obtain statistics on these care cen-

ters, we use the SAE database. There are more than 50 tables in the SAE. Only four tables were necessary. We start with the table “filtre” (n=4,041 hospitals) that gathers the general information about the hospital and the list of services it has. Then we use the “mco” table (n=1,650 hospitals) which contains statistics on care centers with medical surgery or obstetric activity. The table “cancero” (997 hospitals) gathers statistics about oncology activity. Finally, the table “blocs” (1,057 hospitals) gathers information about surgery room activity. We merge the care centers dataset extracted from the PMSI with the SAE tables. “finess”, “filtre” and “mco” are merged with an inner join. This operation will remove care centers that do not declare MCO activity in the SAE. The tables “cancero” and “blocs” are merged with a left join, so that care center with no oncology or surgery activity could remain in the dataset. The missing values were filled with 0. The final merged dataset has 1,588 care centers.

Metropolitan France is divided into 13 regions, 96 departments, and around 35,000 municipalities. The number of municipalities changes each year but is roughly stable. Statistics on municipalities are publicly available on various governmental open data platforms. Municipalities and their census statistics are extracted from the Institut national de la statistique et des etudes economiques (INSEE) website. The most up to date data was released in 2021: population data is from 2017 and 2012, socio-demographic data is from 2018. Municipalities latitude and longitude coordinates are retrieved from La Poste open data platform. In the PMSI database, municipalities with small population are merged into “geographic codes”, an aggregation of one or more municipalities. The list of the geographic codes and the municipalities they are linked with are retrieved from the PMSI database. We merge the INSEE dataset with coordinates extracted from La Poste and the geographic codes correspondence. After merging these tables, the final dataset comprises 13 regions, 96 departments, 34,877 municipalities and 5,608 geographic codes. Figure 2.1 summarizes the data sources we cited previously.

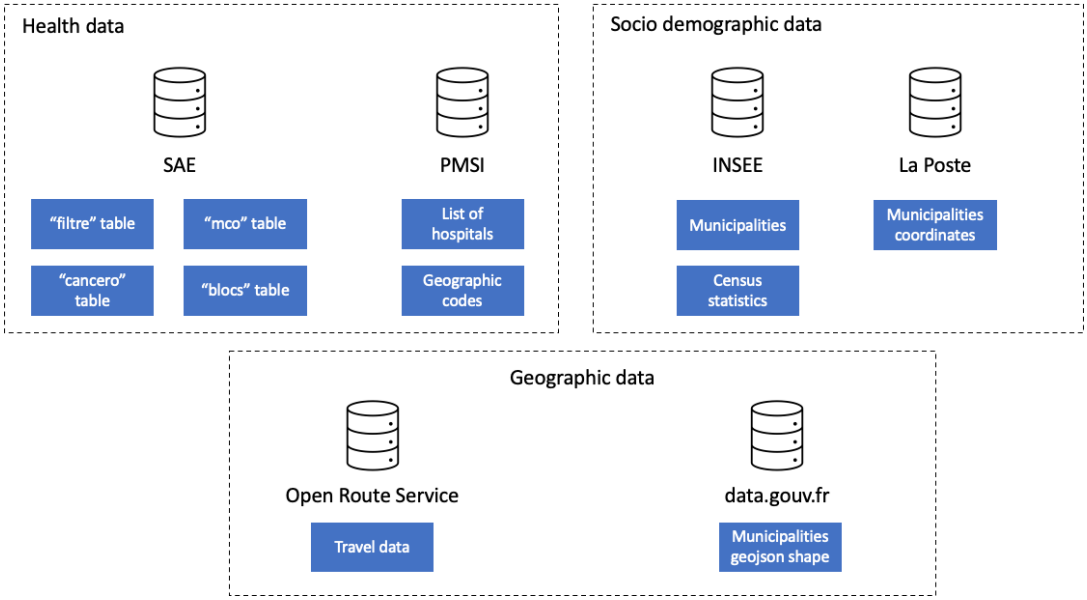


Figure 2.1: Data sources used to characterize the hospitals. We retrieved health data from the Statistiques Annuelles des Etablissements (SAE) and the Programme de Medicalisation des Systemes d'Information (PMSI) databases to characterize the care centers. Then, geographical and socio-demographic data was downloaded from the Institut national de la statistique et des etudes economiques (INSEE) open data platform.

Variable selection

After the previous merge on the SAE health data, we had more than 200 variables for every care center. We selected a list of 24 variables with the help of medical experts. The list of variables and their description are listed in Table 2.1. The variables are either binary when they encode the presence or absence of a service; or continuous when they encode the number of stays. Among the included variables were the number of medical, surgery and obstetric stays; the radiotherapy, chemotherapy, cancer surgery activity and whether the hospital had a dedicated oncology service; the presence of services like palliative care, chronic pain, intensive care, chronic pain; the number of beds; the number of operating rooms. Even though the "cancero" table gives us the number of stays related to oncology, we created a

new variable to encode the oncology activity of a care center. Indeed, the number of stays for radiotherapy or chemotherapy is usually much higher than the number or surgery stays, resulting in an over-representation of these activities compared to surgery. The "cancero" table gives us the number of patients and the number of stays with radiotherapy and chemotherapy per care center. We subtracted the number of radiotherapy and chemotherapy stays from the number of oncology stays. We named this variable "cancero_nb_stays_chirmed". Then we added to this the number of chemotherapy and radiotherapy patients, resulting in a new variable that we refer as "oncology_activity". Finally, log transformation is applied to continuous data and standard scaling (0 mean and unit variance) on every variable.

SAE table	**Variable name**	**Variable definition**	**Distribution**
filtre	chirambu	Outpatient surgery activity	Binary
filtre	chimio	Chemotherapy activity	Binary
filtre	rth	Radiotherapy activity	Binary
filtre	bloc	Surgery activity	Binary
filtre	bio	Medical biology or anatomopathological activity	Binary
filtre	rea	Intensive care unit	Binary
filtre	medic	Medication circuit	Binary
filtre	douleur	Chronic pain	Binary
filtre	palia	Palliative care	Binary
filtre	chircancer	Cancer surgery	Binary
MCO	sejhc_med	Number of inpatient medical stays	Continuous
MCO	sejhc_chi	Number of inpatient surgery stays	Continuous
MCO	sejhp_med	Number of outpatient medical stays	Continuous
MCO	sejhp_chi	Number of outpatient surgery stays	Continuous
MCO	lit_mco	Number of MCO beds	Continuous
blocs	salchir	Number of surgery operating rooms	Continuous
blocs	salambu	Operating rooms dedicated to outpatient surgery	Continuous
cancero	cancero_A1	Use chemotherapy for cancer treatment	Binary
cancero	cancero_A2	Use radiotherapy for cancer treatment	Binary
cancero	cancero_A3	Has an oncology dedicated unit	Binary
cancero	cancero_A11	Number of patients treated with chemotherapy	Continuous
cancero	cancero_A17	Number of patients treated with radiotherapy	Continuous
-	cancero_nb_stays_chirmed	Number of oncology medical or surgery stays	Continuous
-	cancero_activity	Oncology activity	Continuous

Table 2.1: List of the variables used for clustering, and their definitions. All the variables except cancero_nb_stays_chirmed and cancero_activity are coming from SAE. The variables are either binary or continuous. Oncology activity is the sum of cancero_nb_stays_chirmed, cancero_17 and cancero_A11.

Principal Component Analysis (PCA)

PCA is dimensionality-reduction method. It is used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one. The new dataset still contains most of the information in the large set. Dimensionality reduction trades accuracy for simplicity and has multiple ad- vantages. First, dimensionality reduction removes redundant and highly correlated features. Then training statistical models on reduced data is easier and less computationally expensive. Moreover, dimensionality reduction makes it possible to visualize large dimensional data. In practice, PCA projects the original data onto new directions, referred as components. Each component explains some of the variance from the original dataset. Keeping the n components with maximum variance and dropping the other ones performs the actual dimensionality reduction. We call "explained variance" the sum of the variance explained by the components kept. PCA is relatively easy to interpret, as each component is a linear combination of the input variables. The contributions of each input variable to the PCA components are called loading scores. We apply the PCA algorithm to the SAE dataset that describes the care centers. We used Python's scikit-learn [137] implementation of the PCA, since it's very well documented and maintained. The input data has 24 variables, and we perform a dimensionality reduction with $n = 2$ components, explaining 63% of the total variance. We tried different number of components, from 2 to 5, but we found 2 gave good and easy to interpret results.

2.1.2 Clustering

Clustering is the task of grouping data points in such a way that points in the same group are closer to each other than to those in other groups. It is an unsupervised Machine Learning algorithm and does not need labelled data to train on. There are different types of clustering methods and different algorithms. Hard clustering is when each point belongs to a cluster or not. Soft clustering is when each point belongs to each cluster to a certain degree. There are many clustering algorithms, surveyed in [138]. We want to run a clustering algorithm on the PCA reduced dataset to automatically isolate care centers with similar statistics. We tried

several algorithms, and, in our case, Spectral Clustering [139] worked best.

Spectral Clustering [139] helps us overcome two major problems in clustering: one being the shape of the cluster and the other is determining the cluster centroid. K-means algorithm generally assumes that the clusters are spherical or round. In spectral, the clusters do not follow a fixed shape or pattern. We now explain more formally how spectral clustering works. Consider a set of data points $x_1, ..., x_n$ and some notion of similarity $s_{ij} \geq 0$ between all pairs of data points x_i and x_j. The intuitive goal of clustering is to divide the data points into several groups such that points in the same group are similar and points in different groups are dissimilar to each other. We represent the data is in form of the similarity graph $G = (V, E)$, where each vertex v_i in this graph corresponds to a data point x_i. Two vertices x_i and x_j are connected if the similarity s_{ij} between them is positive or larger than a certain threshold, and the edge is weighted by s_ij. The problem of clustering can be reformulated as such: find a partition of the graph such that the edges between different groups have very low weights, and the edges within a group have high weights. The input of the spectral clustering algorithm are the similarity matrix $S \in \mathbb{R}^{n \times n}$ and the number of clusters k to construct. From the similarity matrix, we compute the weighted adjacency matrix $W = (w_{ij})_{i,j=1,...,n}$, where w_{ij} is the weight carried by the edge between two vertices x_i and x_j. If the two vertices are not connected, $w_{ij} = 0$. The degree d_i of a vertex $v_i \in V$ is defined as the sum of all its related weights w_{ij}. The degree matrix D is the diagonal matrix with degrees $d_1, ..., d_n$ on the diagonal. From the W and D matrices, we compute the unnormalized Laplacian $L = D - W$. Then, we compute the first k eigenvectors $u_1, ..., u_k$ of L, and let $U \in \mathbb{R}^{n \times k}$ be the matrix containing the eigenvectors as columns. Then, let $y_i \in \mathbb{R}^k$ be the vector corresponding to the i-th row of U. Cluster the points $(y_i)_{i=1,...,n}$ with the k-means algorithm into clusters $C_1, ..., C_k$. Again, we used Python's scikit-learn Machine Learning library [137], which implemented the spectral clustering algorithm. The parameters were left as default. Hence, the affinity matrix was computed using a radial basis function kernel: $exp(d(X, X)^2)$ with X the input matrix and $d(X, X)$ the euclidean distance. Regarding the number k of clusters, we tried various values from 2 to 10 and manually interpreted the

results with medical experts. We found that 8 clusters gave the most interpretable groups.

2.1.3 Grouping hospitals based on their collaborations

We are now interested in clustering the hospitals based on patients transfers. We call co-occurrence between two hospitals the number of patients that visited these two hospitals during its care pathway. The larger the co-occurrence number is, the more collaboration there is between the two hospitals. The diagram on Figure 2.2 illustrates the co-occurrence definition.

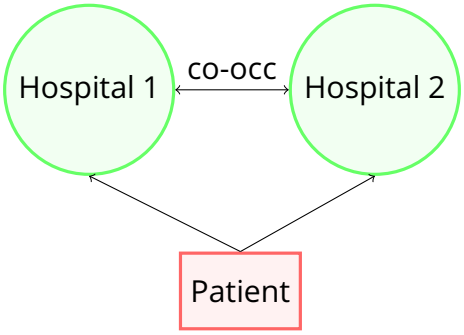


Figure 2.2: Co-occurrence diagram between two hospitals. When a single patient visits two hospitals 1 and 2 during its pathway, we count a co-occurrence between these hospitals. The more patients visit two distinct hospitals, the stronger the co-occurrence link is.

We model the hospitals and their collaborations as a graph, where the nodes are the hospitals, and the edges are weighted by the number of co-occurrences between the hospitals. The task we wish to achieve is community detection over this graph. Intuitively, we seek to find communities of hospitals that frequently interact together by exchanging patients, as illustrated on Figure 2.3.

Community detection algorithms are used to evaluate how groups of nodes are clustered or partitioned together. Detecting communities on graphs is a very active research topic, with many concrete applications [140]. There are different ways to perform community detection on a graph [141]. Several approaches have been developed to learn latent node representations based on the graph topology. These latent representations encode

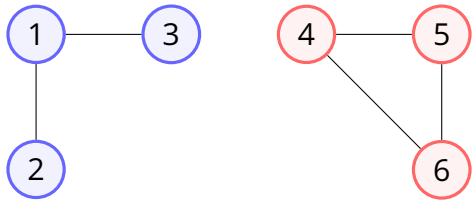


Figure 2.3: Co-occurrence graph between the hospitals. Hospitals are represented as nodes, and edges are weighted by the number of co-occurrences between them. On this graph, there are two communities colored in blue and red that we would like to retrieve.

the graph structure in a continuous vector space, that can be exploited by statistical models [142]. Learning graph representations was traditionally performed with Laplacian regularization. However, research shifted towards learning graph embeddings [143], inspired by the skip-gram model [144]. With such approaches, node embeddings are learned so that nodes that are strongly connected are close in the latent space. Once the embeddings are learned, common statistical learning tasks can be performed such as graph classification, link prediction, or community detection [141]. Recently, the Variational Graph Auto-Encoder (VGAE) was introduced to learn latent representations for undirected graphs, in an unsupervised manner [145]. This framework is based on the Variational Auto Encoder model [146], with a Graph Convolution Network (GCN) [143] encoder and an inner-product decoder. GCN is similar to a regular Convolution Layer used mostly in computer vision. In computer vision, the input neurons are multiplied with a set of weights that are commonly known as filters or kernels. The filters act as a sliding window across the whole image and enable to learn higher level features from the neighboring cells. In GCN, the filters are moved across the graph nodes, and learn features from the neighboring nodes. The hidden representation of a given node can be obtained as the average value of the current node features along with its neighbors. Based on the learned representations of every node, the inner-product decoder aims at reconstructing the adjacency matrix of the input graph. That way, the network will learn similar latent vectors for nodes that are strongly connected in the graph. One advantage of this method is that we can use node features to learn the representations.

In our case, we use the co-occurrence network between the n hospitals as input graph.

This is a non directed weighted graph, where strongly connected nodes are hospitals with many co-occurrences. We ran the VGAE model over the adjacency matrix of the graph, without using nodes features. We a chose a latent representation of size $k = 32$. The output was a matrix $Z \in \mathbb{R}^{n \times k}$ corresponding to the embedding vectors of each hospital node. We ran a TSNE [147] dimensionality reduction algorithm on top of Z to obtain a 2D representation of every hospital. Finally, we performed a clustering on top of the reduced data, with the DBSCAN algorithm [148]. DBSCAN stands for "Density-based spatial clustering of applications with noise". The algorithm can discover clusters of different shapes and sizes, which might contains noise and outliers. The algorithm takes two parameters: the minimum number of points n to form a cluster from; and a distance measure ϵ to locate points within each other. For every point in the dataset, if there are at least n points within a radius ϵ, assign them to the same cluster. The clusters are then expanded recursively by repeating this step for all the remaining points. One of the advantage from this algorithm is that we do not need to specify the number of expected clusters. However, the parameters are sometimes hard to tune, and good initial values should be set with care.

2.2 Results

We first describe the spatial distribution and specificities of the 1,662 hospitals included in this study. There are different types of hospitals in France: Centre Hospitalier (CH) (n=667) and Centre Hospitalier Regional / Universitaire (CHR/U) (n=142) are state-run hospitals; Centre de Lutte Contre le Cancer (CLCC) (n=26) and Participant au Service Public Hospitalier (PSPH) or Etablissement a But Non Lucratif (EBNL) (n=142) are both private hospitals of collective interest, though CLCC are oncology dedicated; private hospitals (n=606) are privately run and for-profit. The non Medecine, Chirurgie, Obstetrique (MCO) care centers with radiotherapy activity (n=79) are mostly private practice structures and are referred as Other. Table 2.2 shows the number of care centers and their oncology activity per hospital type and region. Most of the care centers are public, but a non-insignificant part are private. CLCC

represent only 1.6% of the care centers, yet they are responsible for 14.2% of the overall oncology activity. The care centers are unevenly distributed across the country. For instance, Corse and Centre-Val-de-Loire are the only two regions with no CLCC care centers. Moreover, the proportion of oncology activity per hospital type varies from a region to another. For instance, in Nouvelle-Aquitaine, 47.1% of the oncology activity is handled by private care centers, whereas in Provence-Alpes-Cote-d'Azur it is 21.4%.

Variable value per region		**Hospital Type**						
N = number of centers A = oncology activity (radio. + chemo. + surgery)		CH (n=667)	CH (n=142)	CLCC (n=26)	Other (n=79)	PSPH/EBNL (n=142)	Privé (n=606)	All n=1,662
Auvergne-Rhône-Alpes	N	98 (49,2%)	21 (10,6%)	2 (1%)	7 (3,5%)	13 (6,5%)	58 (29,1%)	199
	A	34,597 (26,7%)	31,706 (24,5%)	16,966 (13,1%)	6,710 (5,2%)	6,146 (4,7%)	33,297 (25,7%)	129.422
Bourgogne-Franche-Comté	N	53 (64,6%)	4 (4,9%)	1 (1,2%)	5 (6,1%)	2 (2,4%)	17 (20,7%)	82
	A	12,238 (27,6%)	10,621 (24%)	5,844 (13,2%)	4,405 (9,9%)	657 (1,5%)	10,571 (23,8%)	44.336
Bretagne	N	38 (33%)	8 (7%)	1 (0,9%)	6 (5,2%)	11 (9,6%)	51 (44,3%)	115
	A	15,953 (27%)	11,020 (18,6%)	6,341 (10,7%)	5,553 (9,4%)	2,050 (3,5%)	18,199 (30,8%)	59.116
Centre-Val de Loire	N	29 (46,8%)	4 (6,5%)	0 (0%)	6 (9,7%)	2 (3,2%)	21 (33,9%)	62
	A	6,989 (19,6%)	11,524 (32,2%)	0 (0%)	5,137 (14,4%)	32 (0,1%)	12,058 (33,7%)	35.74
Corse	N	7 (53,8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (46,2%)	13
	A	3,486 (66,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,773 (33,7%)	5.259
Grand Est	N	70 (41,7%)	17 (10,1%)	3 (1,8%)	6 (3,6%)	30 (17,9%)	42 (25%)	168
	A	17,428 (19,6%)	22,123 (24,9%)	13,176 (14,8%)	6,793 (7,7%)	7,683 (8,7%)	21,553 (24,3%)	88.756
Hauts-de-France	N	56 (40%)	11 (7,9%)	1 (0,7%)	11 (7,9%)	12 (8,6%)	49 (35%)	140
	A	21,864 (26%)	15,934 (19%)	6,947 (8,3%)	8,618 (10,3%)	5,242 (6,2%)	25,399 (30,2%)	84.004
Île-de-France	N	40 (47,6%)	5 (6%)	4 (4,8%)	6 (7,1%)	3 (3,6%)	26 (31%)	84
	A	7,573 (14,9%)	7,947 (15,7%)	14,210 (28%)	5,419 (10,7%)	0 (0%)	15,627 (30,8%)	50.776
Normandie	N	70 (44,9%)	10 (6,4%)	1 (0,6%)	7 (4,5%)	12 (7,7%)	56 (35,9%)	156
	A	37,844 (33%)	26,244 (22,9%)	7,477 (6,5%)	7,157 (6,2%)	2,824 (2,5%)	33,271 (29%)	114.817
Nouvelle-Aquitaine	N	66 (37,7%)	14 (8%)	2 (1,1%)	7 (4%)	6 (3,4%)	80 (45,7%)	175
	A	14,735 (12,1%)	20,915 (17,2%)	16,047 (13,2%)	11,572 (9,5%)	1,098 (0,9%)	57,374 (47,1%)	121.741
Occitanie	N	34 (44,7%)	5 (6,6%)	3 (3,9%)	4 (5,3%)	5 (6,6%)	25 (32,9%)	76
	A	11,901 (18,9%)	11,374 (18,1%)	12,564 (19,9%)	3,422 (5,4%)	3,916 (6,2%)	19,822 (31,5%)	62.999
Pays de la Loire	N	53 (34,4%)	10 (6,5%)	3 (1,9%)	4 (2,6%)	15 (9,7%)	69 (44,8%)	154
	A	14,632 (13,6%)	16,533 (15,4%)	21,924 (20,4%)	6,172 (5,7%)	10,918 (10,2%)	37,176 (34,6%)	107.355
Provence-Alpes-Côte d'Azur	N	53 (22,3%)	33 (13,9%)	5 (2,1%)	10 (4,2%)	31 (13%)	106 (44,5%)	238
	A	24,390 (12,6%)	66,406 (34,2%)	34,028 (17,5%)	12,817 (6,6%)	14,981 (7,7%)	41,577 (21,4%)	194.199
Grand Total	N	667 (40,1%)	142 (8,5%)	26 (1,6%)	79 (4,8%)	142 (8,5%)	606 (36,5%)	1662
	A	223,630 (20,4%)	252,347 (23%)	155,524 (14,2%)	83,775 (7,6%)	55,547 (5,1%)	32,7697 (29,8%)	1,098,520

Table 2.2: Number of care centers (N) and overall oncology activity (A) per hospital type and region. Oncology activity is the sum of the number of patients with radiotherapy or chemotherapy, and the number of medical or surgery stays related to cancer. CH and CHR/U are public hospitals; CLCC and PSPH/EBNL are private hospitals of collective interest, though CLCC are oncology dedicated; private hospitals are for-profit. Other hospitals are mostly private practice radiotherapy structures. The percentages sum to 100% row-wise. In Nouvelle-Aquitaine, 47.1% of the oncology activity is handled by private care centers, whereas in Provence-Alpes-Cote-d'Azur it is 21.4%.

2.2.1 Oncology specialization label

While it is obvious that CLCC care centers are suited for oncology care, it is difficult to assess the degree of oncology specialization for other care centers. Our clustering algorithm assigned the n=1,662 care centers into 8 clusters, sorted by oncology specialization. The PCA and clustering results are visible on Figure 2.4. The scatter plots (A) and (B) display the hospitals as points in the 2-dimensional PCA space, colored by assigned cluster (A) and hospital category (B). We see two main groups of points on this scatter plot, one on top and one on the bottom. These points are well separated along the second PCA component. From plot (C), we can interpret the PCA components. The first one is correlated with almost every input variable, meaning that the higher this component is, the larger and the more developed the hospital is. Regarding the second component, it is correlated with oncology dedicated variables, especially radiotherapy. This means that hospitals with a large value along the second component are dedicated to oncology and have a radiotherapy activity. From this, we understand that points with large values along the two components are large hospitals with an important oncology activity. This seems to be the case for hospitals in clusters 1 and 2 (A). When we look at the hospitals categories on plot (B), we notice that these points on the top right side of the figure are essentially CLCC, which makes sense since these hospitals are fully dedicated to oncology by design. However, there are also hospitals from all the other categories, which would have been less easy to identify as oncology experts.

The Figure 2.5 shows the distribution of some of the key health services per cluster. These services are biology, radiotherapy, chemotherapy, cancer surgery, intensive unit, palliative care, oncology unit, medication circuit, surgery, and outpatient surgery. The three oncology services are cancer surgery, radiotherapy, and chemotherapy. We see that care centers from clusters 1 (n=79) and 2 (n=39) all have these 3 services, hence they are the most suited hospitals for oncology care. Centers from cluster 3 (n=451) have cancer surgery and chemotherapy but lack radiotherapy. The most part of the n=381 centers from cluster 4 have cancer surgery, but no radiotherapy nor chemotherapy. Care centers from cluster 5 (n=2) and cluster 6 (n=7) have radiotherapy and chemotherapy services, but no cancer

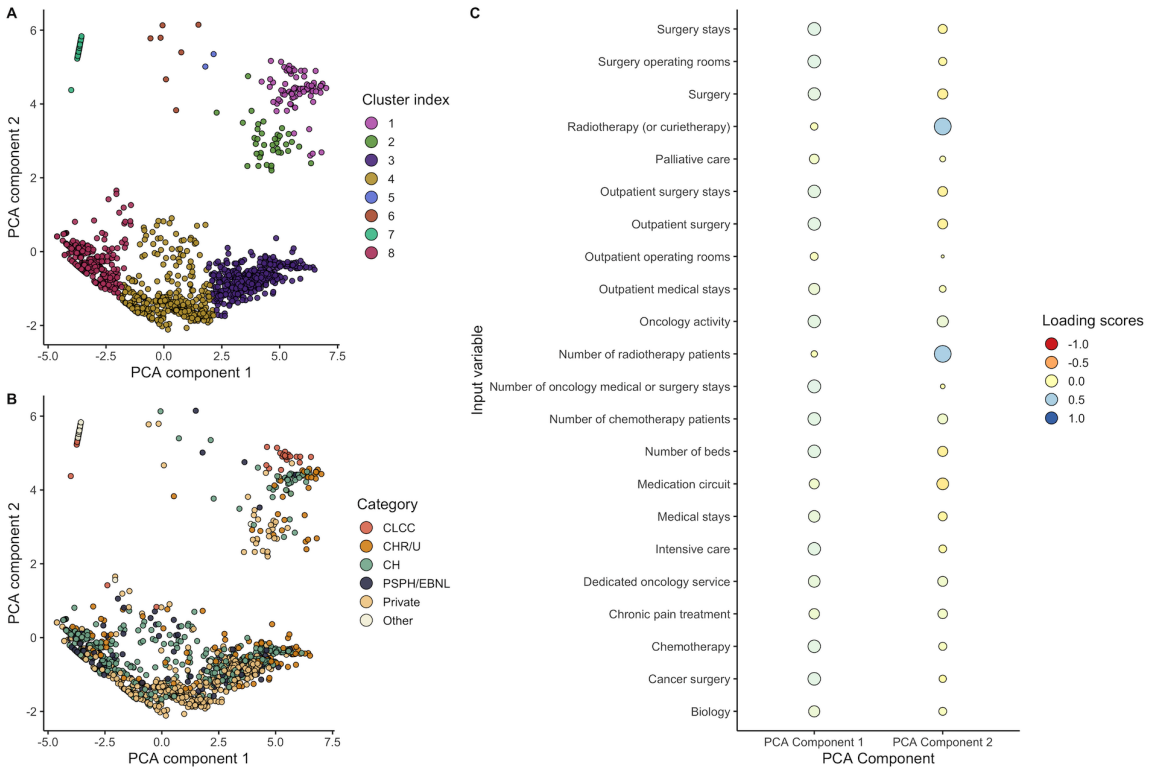


Figure 2.4: PCA interpretation. Care centers are showed as points in the 2-dimensional PCA space. Points are colored by cluster index (A) and hospital type (B). CLCC care centers are close together in the PCA space, proving they have similar activity and services distribution. PCA components are a linear combination of the input variables (C). The loading scores reflect how much the input variable contributed to the PCA component. Component 1 is associated with most of the variables, while component 2 is linked with radiotherapy variables. Hence, we interpret component 1 as hospital size and component 2 as oncology specialization.

surgery. Care centers in cluster 7 (n=77) are dedicated to radiotherapy and mostly private practice structures. Finally, care centers 8 (n=626) have none of the 3 oncology services. To sum up, hospitals from clusters 1 and 2 (n=118) are all-in-one care centers that provide the most ideal oncology care. Centers from clusters 3 and 4 (n=382) provide oncology care but will have to be coordinated with additional structures during the pathways. Hospitals within clusters 5, 6 and 7 (n=86) are not allowed to perform cancer surgery but provide chemother-

apy or radiotherapy. The remaining n=626 care centers in cluster 8 are not equipped for oncology care.

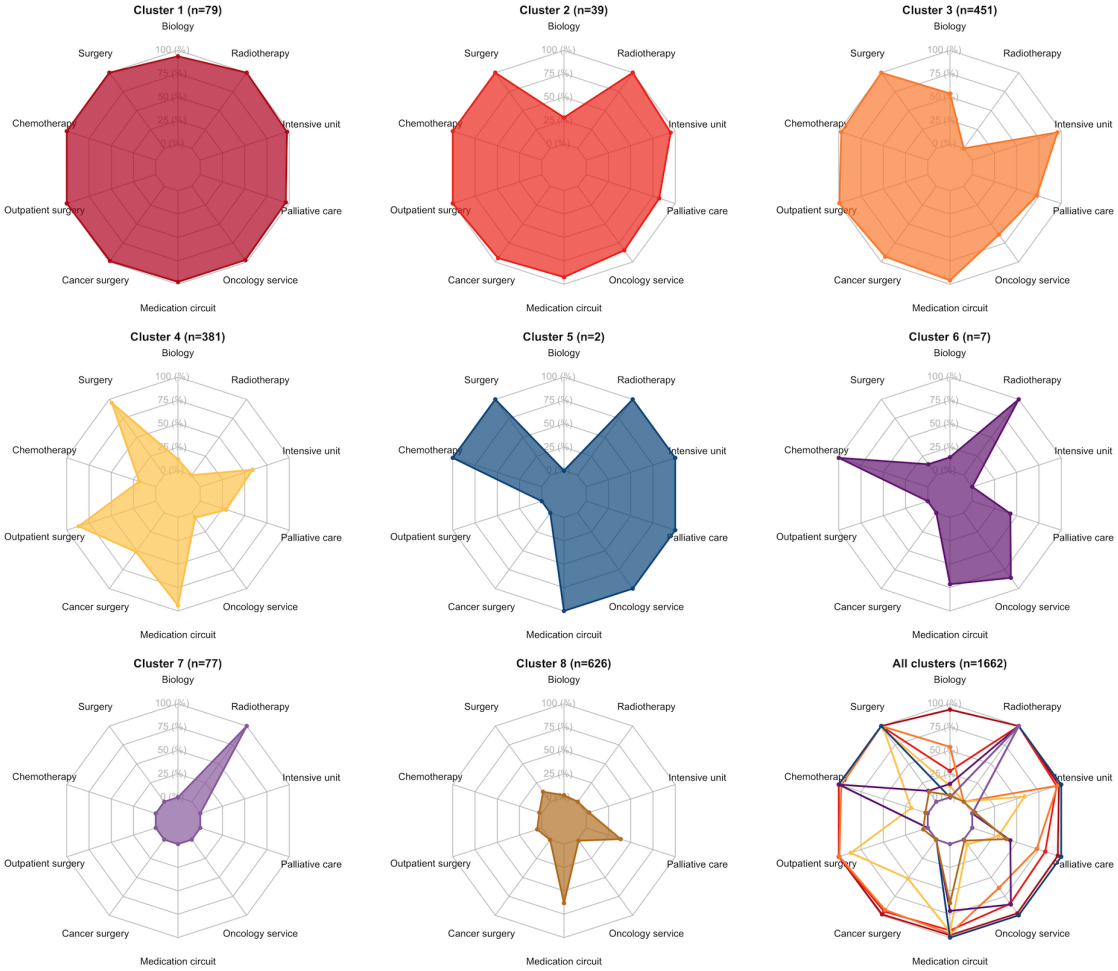


Figure 2.5: Distribution of the care centers services and equipment per cluster. Each radar plot axis shows the percentage of the care centers within the cluster that have the corresponding attribute. In Cluster 1, the care centers have all the listed services. In cluster 8, the centers have almost none of the services. Care centers from cluster 1 (n=79) and cluster 2 (n=39) are the most suited for oncology care.

We now look at the hospital categories distribution among each cluster. Hospital types are unevenly distributed among the clusters as illustrated on Figure 2.6. For instance, 76.9% of the CLCC care centers are placed in cluster 1, as they are the most specialized centers.

In cluster 7, we find external radiotherapy units of some CLCC centers, and private practice structures, classified as “Other”. The proportion of private care centers varies as well: cluster 1 has almost no private care center while cluster 2 has 61.5% of private hospitals. Thus, it appears that many privately held hospitals do not have a biology activity, but they have all the other services. This illustrates that the hospital category could not be used to assess the oncology specialization, apart from the CLCC category.

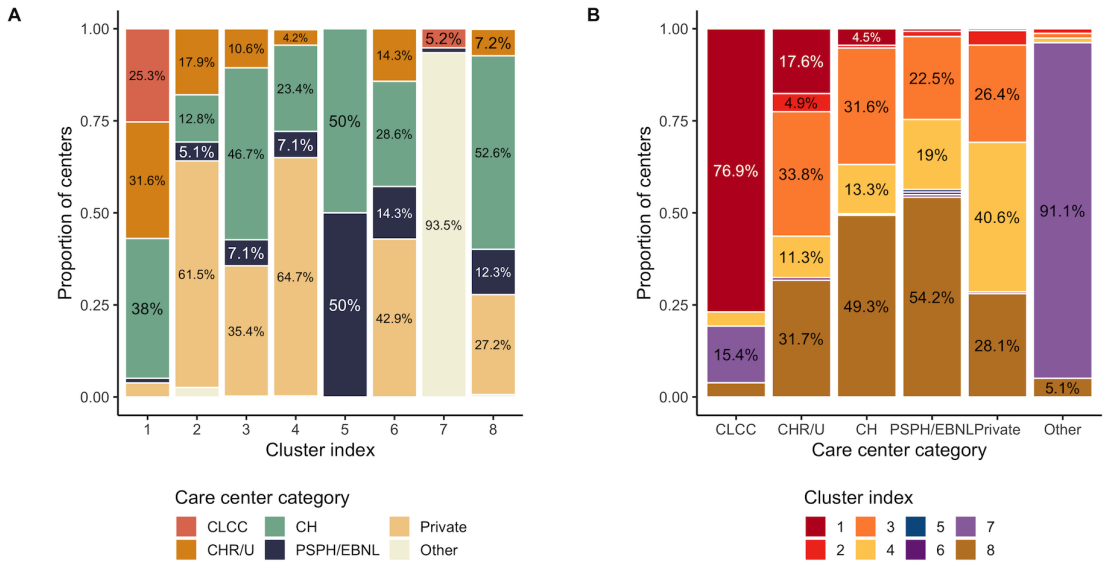


Figure 2.6: Comparison between hospital types and assigned clusters. The majority of the CLCC care centers are grouped together in cluster 1. Moreover, cluster 1 has a very low percentage of private hospitals, whereas this proportion is the much higher in cluster 2. “Other” care centers are mostly private practice radiotherapy structures, and they are regrouped in cluster 7.

The Figure 2.7 shows the percentage of oncology activity covered by each cluster. Most of the oncology activity is handled by care centers from clusters 1 and 3. The activity of the 79 hospitals in the cluster 1 combined equals 36.8% of the total activity. This is almost as large as the activity of the n=451 hospitals from cluster 3. As mentioned in the centralization of care benefits, the care centers from the cluster 1 probably have the largest expertise due to the large volume of cancer patients they treat.

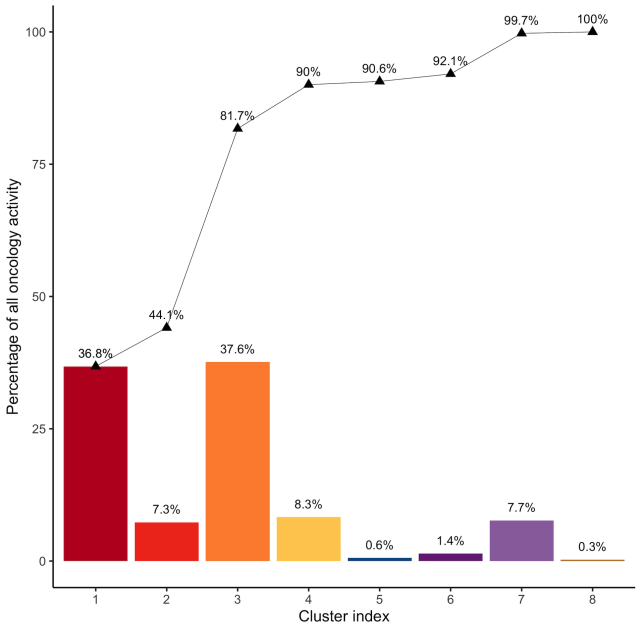


Figure 2.7: Cumulative sum of the oncology activity, per cluster. Most of the oncology activity is handled by care centers from clusters 1 and 3. While there are only n=79 care centers in cluster 1, their total activity is almost as large as the n=451 care centers from cluster 3.

Finally, we study the spatial distribution of the hospitals in metropolitan France, and compare it with the population density distribution. In 2018, the population in France was 66,993 million. Mainland France hosts 64,812 million inhabitants (96.8%), while the remaining 2,181 million (3.2%) live in overseas departments and regions . Metropolitan France is divided into 13 administrative regions and 96 departments. The population density in France is unevenly distributed . In 2020, the overall population density in metropolitan France was 119 inhabitants per square kilometer. Ile-de-France region has the highest population density with 1,022 inhabitants per square kilometer. Density in other regions in metropolitan France range between 40 and 187 inhabitants/km^2. Denser areas are located near the coastline and around the largest cities like Paris, Marseille, Lyon, Strasbourg, Toulouse, or Bordeaux. The middle of the country is rural, and the population densities are low. While there are a great variety of regions and landscapes, the country is becoming more urbanized. This “ru-

ral exodus" is largely responsible of what is known as the "empty diagonal", a band of very low-density population that stretches from the southwest to the northeast. On Figure 2.8, Map (A) shows the metropolitan France map, with municipalities colored by population density cuts. The various bins are: <30; 30-50; 50-100; 100-200; and >200 inhabitants per km^2. The hospitals are displayed as pictograms, sized by oncology activity and colored by their assigned cluster. Unsurprisingly, the largest hospitals and the most specialized in oncology are located in densely populated areas. The box plot (B) shows the population distribution of the municipalities where the hospitals are located, by cluster index. As expected, the municipalities where hospitals from the cluster 1 are more populated. Bar plots (C) and (D) show the number of hospitals by cluster index for every region. Plot (C) shows the absolute number where plot (D) shows the number of hospitals per 100,000 inhabitants.

2.2.2 Collaborations between hospitals

We now describe our results on collaborations between hospitals. We selected patients with a cancer diagnosis during the year 2018, regardless of the cancer type. These patients visited a total of n=1,433 distinct hospitals. We computed the co-occurrence input matrix $X \in \mathbb{R}^{n \times n}$ from this dataset. We counted a co-occurrence between two hospitals every time a patient visited both hospitals during the year. We stress that it did not have to happen during the same stay. We then ran the VGAE model to learn the nodes representations, and performed dimensionality reduction with TSNE. Finally, the DBSCAN algorithm outputs 26 communities, and failed to find a community for 13 hospitals. We displayed the retrieved communities on Figure 2.9. On this map, the hospitals are displayed as pictograms, sized by their oncology activity, and colored by the retrieved community, corresponding to the DBSCAN cluster. The links between the hospitals are the co-occurrences, and we only displayed links with more than 60 co-occurrences for clarity.

We studied the hospitals distribution and geographical spread, as illustrated on Figure 2.10. The barplot (A) displays the number of hospitals per community, and oncology specialization cluster. We notice that most of the communities have hospitals from the most

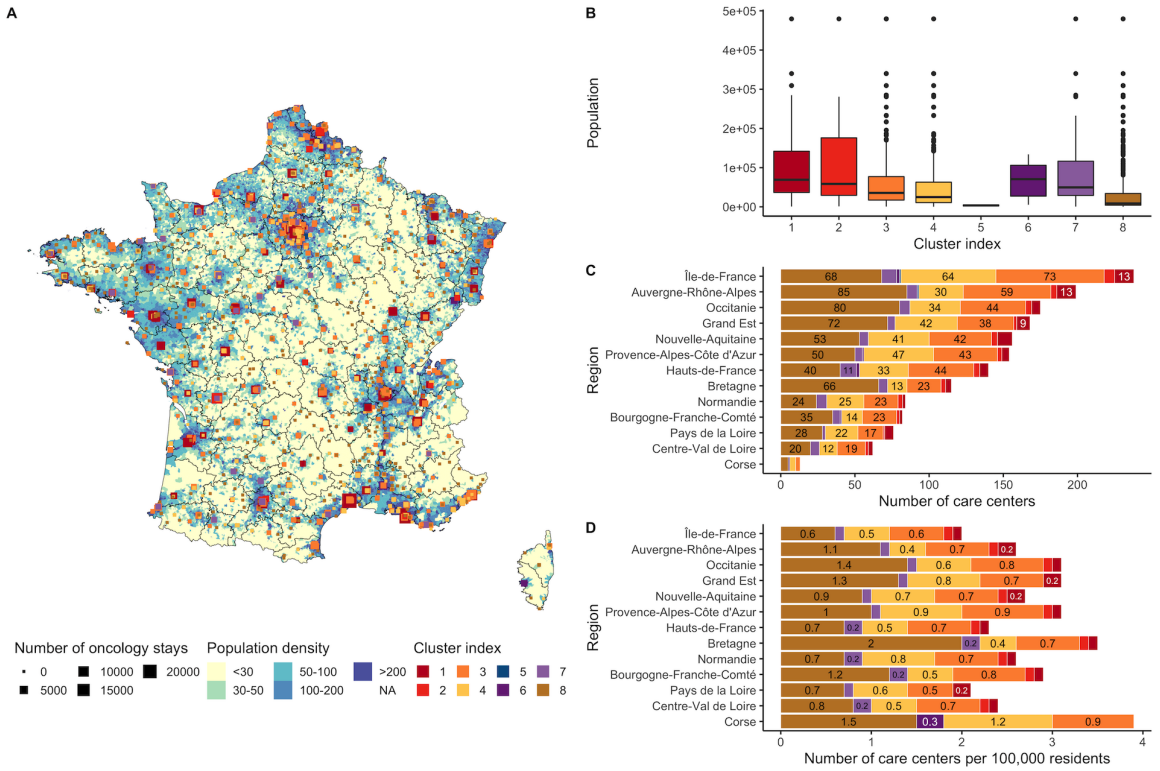


Figure 2.8: Care centers spatial distribution, compared with population density. Population density in metropolitan France is unevenly distributed across the country (A). Areas in the middle, near the Pyrenees and the Alps have very low population densities. The most specialized care centers are in dense areas and in large municipalities (B). While Ile-de-France has the highest number of care centers, it has the least care centers per 100,000 inhabitants.

specialized clusters. The next plot (B) displays the oncology activity per community and hospital cluster. As seen before, most of the activity is handled by hospitals from the most specialized clusters 1, 2 and 3. From this data, we identify some communities that do not have any oncology activity. The community labeled "-1" refers to the hospitals that could not be assigned to any cluster by the DBSCAN algorithm. Finally, we studied the geographical spread of the communities, to assess whether the hospitals are located closely to each other or far apart, as illustrated on plot (C). Intuitively, hospitals from the same communities should be located in the same area so that patients exchanges are easy and do not require long travels. This spread value was computed as the maximum square area containing all

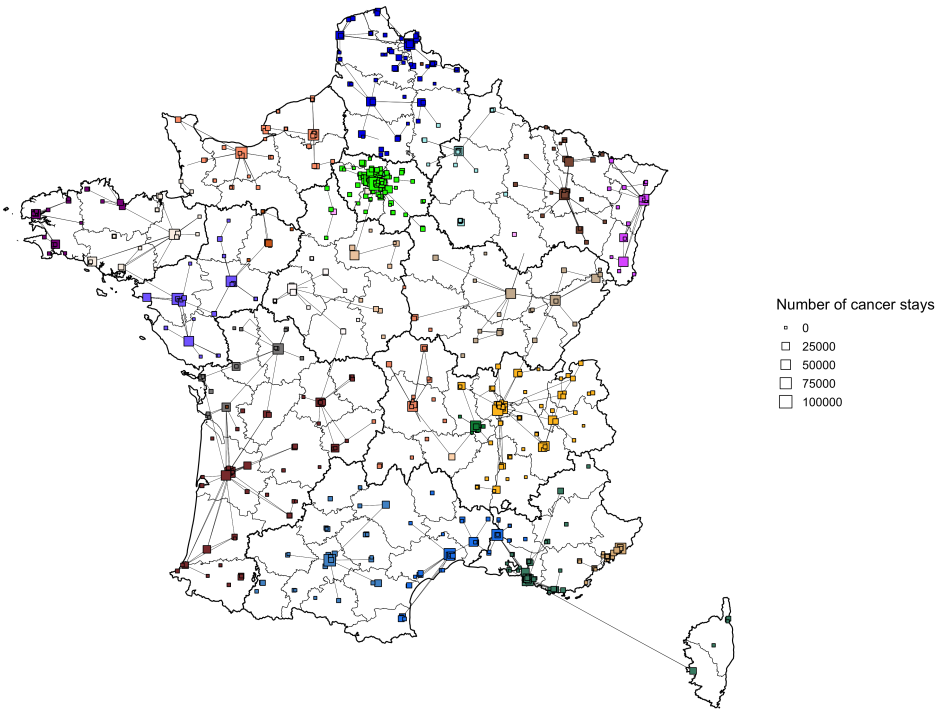


Figure 2.9: Community detection in France, learned on the co-occurrence matrix. The hospitals are displayed as pictograms, sized by their oncology activity, and colored by the retrieved community, corresponding to the DBSCAN cluster. The links between the hospitals are the co-occurrences, and we only displayed links with more than 60 co-occurrences for clarity.

the hospitals from the community, obtained from Equation (2.1).

$$spread = (longitude_{max} - longitude_{min}) * (latitude_{max} - latitude_{min}) \tag{2.1}$$

The communities with the maximum spread are "-1" and "1". We already mentioned the "-1" community as the list of 13 hospitals that the DBSCAN algorithm did not cluster. Regarding the "1" community, it contains 307 hospitals, that do not have any oncology activity. Hence, they are not connected to any hospital by the co-occurrence matrix. These hospitals could be either removed from the algorithm, or identified as hospitals to coordinate with nearby

structures with oncology activity.

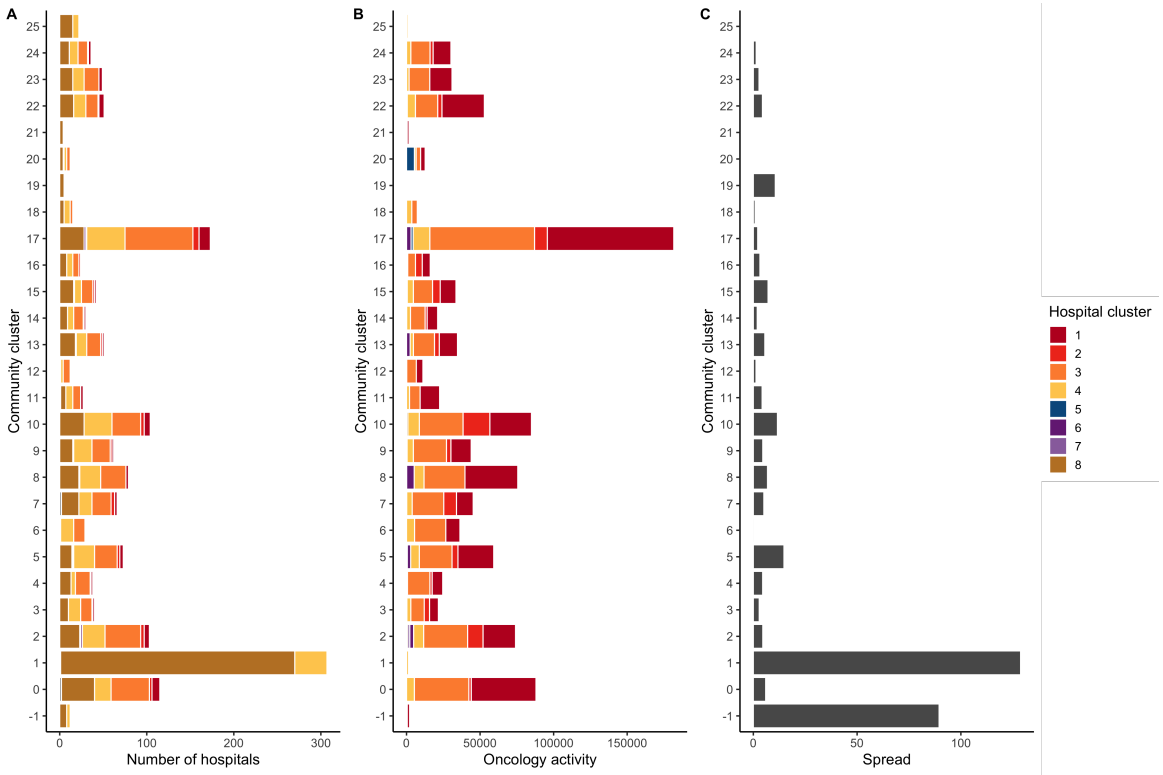


Figure 2.10: Description of the discovered communities. A total of 26 communities were discovered by the DBSCAN algorithm. The barplot (A) displays the number of hospitals per community, and oncology specialization cluster. The next plot (B) displays the oncology activity per community and hospital cluster. Plot (C) illustrates the geographical spread of the communities, to assess whether the hospitals are located closely to each other or far apart.

Finally, we focused on a single community, and picked the one labelled “2”, as illustrated on Figure 2.11. The hospitals are still sized by oncology activity, but they are now colored by the oncology specialization cluster, assigned in Section 2.1.1. From the map displayed on Figure 2.11, we notice that the co-occurrences links are often grouped around the most specialized hospitals, belonging to the cluster 1. We also notice several sub-graphs in this community, again centered around cluster 1 hospitals. By combining the oncology specialization cluster and the collaborations, we can assign the most specialized hospitals as reference hospitals, that could be responsible for coordinating the less specialized structures.

Figure 2.11: Community detection in France, focus on the single community "2". The hospitals are displayed as pictograms, sized by their oncology activity, and colored by the oncology specialization cluster.

2.3 Conclusion

In this chapter, our goal was to characterize all the care centers in France with medicine, surgery or obstetric activity. After gathering health statistics on these hospitals, we ran clustering algorithm to automatically label each center in terms of oncology specialization. The clustering algorithm successfully groups similar hospitals and lets us identify the care centers best suited for oncology care. Some variables in the SAE survey are declarative and potentially differ from the reality. We are aware of this bias, but we do not expect major differences that could distort our clustering results. Receiving treatment in a care center with surgery, chemotherapy and radiotherapy activities is easier for the patient and leads to better care pathways. Care centers from cluster 1 will be the better choice for cancer treat-

ment and correspond to modern oncology care specifications. However, these centers are a minority and sparsely located, essentially in dense areas and in large cities. While the inhabitants of large cities and metropolitan areas will have no problem reaching them, rural areas residents live far away from these centers. This population often has better access to care centers from intermediate clusters. Such centers do not have all the key services and the patients are more likely to visit multiple hospitals during their care pathways. Longer drives to reach a more specialized care center could be considered more acceptable for surgery, where the hospital volume and surgeon expertise matter. However, for more frequent interventions like chemotherapy and radiotherapy especially, patients should prioritize short travels. There is a tradeoff to be found by patients, between care center proximity and care center expertise. This dilemma will be more frequent for patients living in rural areas than patients living in dense cities with large care centers nearby. The different levels of oncology specialization and the uneven spatial distribution of the oncology hospitals should be a reason to improve collaborations between hospitals. If the hospitals with less expertise work closely with oncology dedicated hospitals, the risks for patients to receive inadequate treatment might be lowered. To highlight the currently existing hospitals collaborations, we ran a community detection algorithm on the co-occurrence matrix. The resulting communities are hospitals that frequently share patients together. By crossing the oncology specialization clusters and these communities, we identified patterns of collaborations, where the most oncology specialized hospital was placed at the center of the collaboration network between the smaller hospitals nearby. These new communities could be interpreted as oncology collaboration groups, and they have been defined from real-life patients data. We believe that they could be used to start the reflection around better designed collaboration networks, based on already existing patients exchanges, proximity between hospitals, and complementarity between the hospitals.

Chapter 3

Accessibility to oncology care

This chapter is part of a research article currently under submission. The preprint is available on medrXiv.

3.1 Methods

3.1.1 Spatial Accessibility methods overview

There are several ways to compute accessibility to healthcare as reviewed in [33]. Some methods are very straightforward and as easy as computing ratios per geographical units. Other methods are more sophisticated and can model more real world factors. We detail these methods in the following sections.

Provider-to-population ratios

The easiest and most straightforward SA method is to compute provider-to-population ratios, also referred as supply ratios. The ratio involves some indicator of health service capacity (supply) as numerator; while the denominator is the population size within the area (demand). For instance, when measuring accessibility to primary care, one might use the number of physicians in the area as supply, and area population as demand. The resulting

ratio might be interpreted as the number of physicians per 100,000 inhabitants [149].

Supply ratios are highly interpretable, and relevant for comparisons of supply in large areas. Policy analysts have used these metrics to set minimal standards of supply and identify under-served areas where supply should be increased [149, 150, 151]. However, supply ratios have limitations that often prevent their usage in more detailed analysis. First, they do not account for patient border crossing, which commonly occurs for small areas [152, 153, 154, 155]. Second, supply ratios are constant within the bordered area, which will lead to imprecision and false generalization in large areas. Finally, they do not consider travel impedance, which plays a major role in SA. Consequently the results and interpretations can vary greatly depending on the size, number and configuration of the areal units studied. This problem is well-known to geographers and spatial analysts as the modifiable areal unit problem (MAUP) [156].

Travel impedance to providers

As stated earlier, travel impedance is a key aspect of SA evaluation. It is typically measured from a patient's residence or from the centroid of a population location when the precise location is not available. The impedance can be expressed in different ways: euclidean (straight) distance, travel distance or estimated travel time.

Travel impedance is suited for rural areas, where providers are limited and patients often travel to the nearest choice available. However, travel impedance is less relevant for urban areas. Indeed, there are numerous reasonable options available at a similar distance and patients won't travel to the closest one anymore. Moreover, travel impedance is a poor indicator of availability and should be combined with supply to properly evaluate SA [157].

Gravity models

Gravity models are more sophisticated ways to evaluate SA, based on a modified version of Newton's Law of Gravitation. They were initially developed to predict retail travel [158] and help with land use planning [159]. Gravity models combine both accessibility and availabil-

ity, and work well in urban and rural settings. Gravity models represent the influence of all service points located within a reasonable distance from a population location. The influence is discounted by the increasing distance or travel impedance. The simplest formula for gravity-based accessibility is:

$$A_i = \sum_u \frac{S_u}{d_{iu}^{\beta}} \tag{3.1}$$

In this equation, A_i is the accessibility score at population location i. S_u is the capacity of the service point u, and d_{iu} the travel impedance (e.g. distance or travel time) between population location i and service point u. We set β as a gravity decay coefficient, sometimes referred to as the travel friction coefficient. Intuitively, β represents the change in difficulty of travel as the impedance value increases. The accessibility score increases with higher provider capacity, and decreases with higher travel impedance. Gravity models are an elegant way to compute accessibility, which accounts for border crossing, local variations, and travel impedance. The main drawbacks of this approach is the lack of intuitiveness, and healthcare policy makers prefer to think of SA in terms of provider-to-population ratios or simple distance. Second, it only models supply and does not account for demand. Providers should not be equally accessible if they serve population locations with drastically different population sizes. A proposed solution is to add a population demand factor V_u, to the denominator [160]:

$$V_u = \sum_k \frac{P_k}{d_{ku}^{\beta}} \tag{3.2}$$

Here, P_k is the population size at population location k, and d_{ku} is the distance between the population location k and provider location u. Intuitively, the demand on provider location u is obtained by summing the gravity discounted influence of all population points within a reasonable distance. The improved gravity model is:

$$A_i = \sum_j \frac{S_j}{d_{ij}^{\beta} V_j} \tag{3.3}$$

However, another problem is that the distance decay coefficient, β, is usually unknown and hard to estimate. Its form and magnitude can vary greatly with the service type and population under study [161].

Two Step Floating Catchment Area (2SFCA)

Recently, a new type of method has been developed and is now used in most SA papers. This algorithm is called Two Step Floating Catchment Area (2SFCA) [162]. It is a two-step method that first computes a provider-to-population ratio for each provider location. In the second step, for each population location, an accessibility score is obtained by summing the provider-to-population ratios. For the algorithm to work, a catchment threshold (distance or travel time) must be set. Above this threshold, a provider location is considered unreachable from the population location, and vice versa.

- Step 1: for every provider u, compute its capacity-to-population ratio R_u.
- Step 2: for every population location, compute A_i as the sum all the R_u of the reachable providers.

$$R_u = \frac{S_u}{\sum_{k \in \{d_{ku} \leq d_0\}} P_k} \tag{3.4}$$

$$A_i = \sum_{u \in \{d_{iu} \leq d_0\}} R_u \tag{3.5}$$

The capacity of a provider is balanced by the total population with access to it. A population location that solely has access to low capacities or overcrowded providers will have a low accessibility score. Similarly, a population location will have low accessibility scores if the distance to get to the nearby providers is above the catchment area.

Enhanced Two Step Floating Catchment Area (E2SFCA)

The 2SFCA method does not account for distance decay: a provider is either reachable or not. The E2SFCA [55] addresses this limitation by applying weights to differentiate travel zones in both steps. Consider P_i the population at location i, with $1 \leq i \leq n$ where n is the number of population locations. Similarly, consider S_u the capacity of care center u, with $1 \leq u \leq m$ where m is the number of care centers. Finally, let d_{iu} be the matrix of size $n \times m$ containing the distances between location i and providers u. We consider r sub-catchment zones each associated with a weight W_s, and a distance D_s, with $1 \leq s \leq r$, such that $D_1 < D_2 < ... < D_r$ and $W_1 > W_2 > ... > W_r$. The resulting r travel intervals are $I_1 = [0, D_1], I_2 = [D_1, D_2], ..., I_r = [D_{r-1}-, D_r]$. The accessibility A_i of a population location i is computed in two steps:

- Step 1: for every care center u, compute its weighted capacity-to-population ratio R_u.
- Step 2: for every population location, compute A_i as the sum all the weighted R_u of the reachable providers.

$$R_u = \frac{S_u}{\sum_{s=1}^{r} W_s \sum_{i, d_{iu} \in I_s} P_i} \tag{3.6}$$

$$A_i = \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} R_u \tag{3.7}$$

Multi modal Two Step Floating Catchment Area

The E2SFCA methodology can be enhanced by incorporating both public and private transport modes [163]. The proposed model yields separate accessibility scores for each modal group at each demand point to better reflect the differential accessibility levels experienced by each cohort.

Suppose that each method of travel (car, bus, walking, etc.) necessitates a dedicated transport network and let each such network be referred to as $N_1, N_2, ..., N_M$. In order to

accommodate independent networks for each travel mode into the E2SFCA model, the computation of Step 1 becomes:

$$R_u = \frac{S_u}{\sum_{m=1}^{M} \sum_{s=1}^{r} W_{s,m} \sum_{i,d_{iu,m} \in I_s} P_{i,m}} \tag{3.8}$$

Similarly for Step 2:

$$A_i = \sum_{m=1}^{M} \sum_{s=1}^{r} W_{s,m} \sum_{u,d_{iu} \in I_s} R_u \tag{3.9}$$

Huff model and Two Step Floating Catchment Area

The E2SFCA does not consider competition among multiple healthcare sites available for a population location [164], and therefore it may lead to overestimation for some sites [41]. The Huff Model is a widely accepted method for quantifying the probability of people's selection on a service site out of multiple available ones [165]. It specifically aims to estimate/-model people's choice on a service site with two factors: the distance to the service site; and the attraction of the service site. Let $Prob_{i,j}$ be the probability of population location i visiting service site j, defined by Equation (3.10). In this formula, d_{ij} is the travel time between i and j; β is the distance impedance coefficient, usually set between 1.5 and 2; C_j is the capacity or attractiveness of service site j; and s are the service sites within the catchment D_0 of i.

$$Prob_{i,j} = \frac{C_i d_{ij}^{-\beta}}{\sum_{s \in D_0} C_s d_{is}^{-\beta}} \tag{3.10}$$

Incorporating the $Prob_{i,j}$ term into the E2SFCA steps brings the following equations:

$$R_u = \frac{S_u}{\sum_{s=1}^{r} W_s \sum_{i,d_{iu} \in I_s} Prob_{i,u} P_i} \tag{3.11}$$

$$A_i = \sum_{s=1}^{r} W_s \sum_{u,d_{iu} \in I_s} Prob_{i,u} R_u \tag{3.12}$$

3.1.2 Computing accessibility to oncology care scores

We now explain how we computed our oncology accessibility score. As we want to compute the accessibility to oncology care centers, we chose S_u to be the oncology activity of a hospital u. We define oncology activity as the sum of the number of medical and surgery stays related to cancer, and the number of patients with chemotherapy or radiotherapy. A care center with no oncology activity will have $R_u = 0$ and a municipality that solely has access to this care center u will have $A_i = 0$. We use driving duration as travel impedance metric, and we set the maximum catchment area to a 90-minute drive. In 2018, only 24,152 patients out of 761,057 (3.2%) had travel duration greater than 90 minutes for cancer related pathways. This is low enough to consider that care centers are non-reachable beyond this distance. We divide the catchment area into 3 intervals: $I_1 = (0, 30]$,$I_2 = (30, 60]$ and $I_3 = (60, 90]$. The associated weights are respectively $W_1 = 1$, $W_2 = 0.042$ and $W_3 = 0.09$. These sub catchment areas are set based on the cancer pathways travel duration distributions and validated with medical experts. The weights are the same than the E2SFCA paper [55]. For privacy reasons, municipalities with small populations are grouped in entities called “geographic codes” in the PMSI database. We decided to compute the accessibility score for each geographic code and municipalities that are grouped in the same code will have the same accessibility score.

On Figure 3.1, we display the accessibility score distribution. We compared the results from different methods. The E2SFCA and 2SFCA algorithms were compared, and we used either the oncology activity or overall number of MCO stays as supply variables. As expected, the median accessibility score is much higher when using the MCO stays as supply variable. Using the E2SFCA algorithm rather than the 2SFCA changes the distribution shape, by shifting values to the left, due to the distance weights.

3.2 Results

The oncology accessibility is unevenly distributed across the country, as displayed on Figure 3.2. For better readability, we cut the accessibility scores into 5 quantiles. Q5 colored in

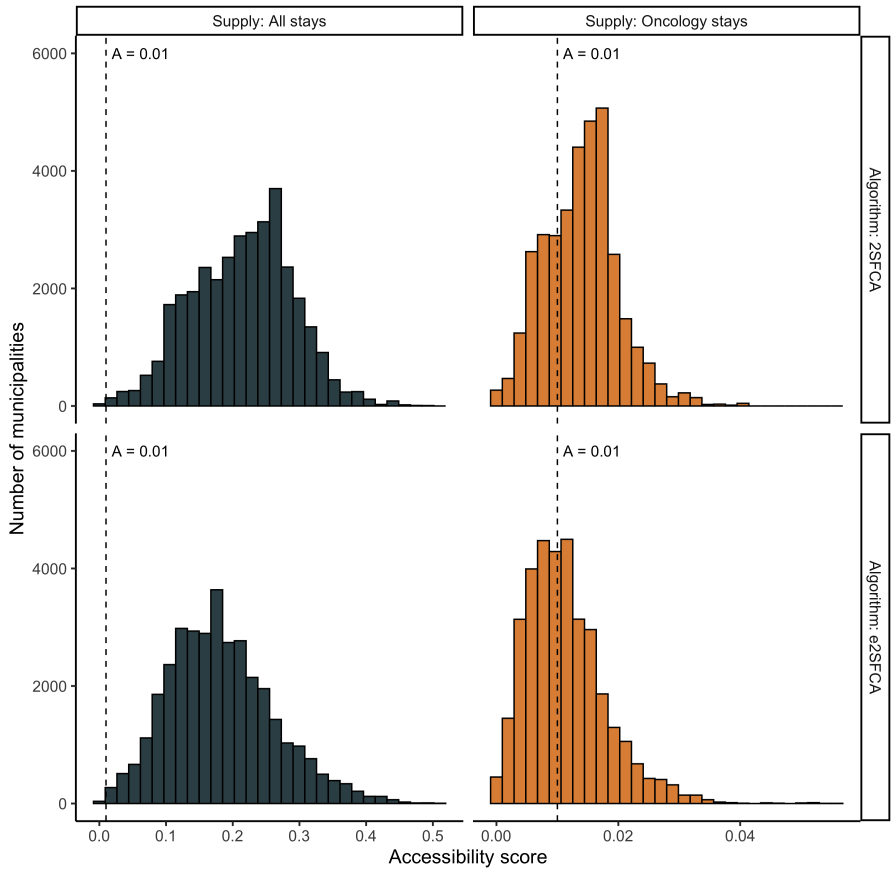


Figure 3.1: Accessibility scores distribution. The accessibility was lower with E2SFCA because of the weight decay. We also studied the influence of the supply variable in the accessibility score. Accessibility is much higher if we use the number of MCO stays as supply, instead of the oncology activity. This makes sense since oncology care centers are less common and the overall MCO activity is higher than the oncology activity.

dark green contains the top 20% accessibility municipalities, and Q1 in light yellow contains the bottom 20% ones. The lowest accessibility zones are mostly located in the center of the country and in mountainous regions like the Alps or the Pyrenees. Plot (B) shows that most of the population (51.6%) lives in top 20% accessibility municipalities, while 6.3% lives in the bottom 20% quantile. On map (A), care centers are displayed as squares, colored by cluster index, and sized by oncology activity. We see that accessibility is highest near the most

specialized care centers. Indeed, the proportion of care centers from specialized clusters decreases in lower accessibility quantiles (C). We then ranked the departments by median accessibility and showed the top-10 and bottom 10 on plot (D). Among the top-5 departments, 4 are in Ile-de-France. Departments from the bottom-10 are rural or mountainous areas like Lozère and Alpes-de-Haute-Provence. We notice disparities within departments as well, as outlined by the large interquartile range in Hérault or Alpes-Maritimes. On the contrary, this spread is very narrow in Ile-de-France departments.

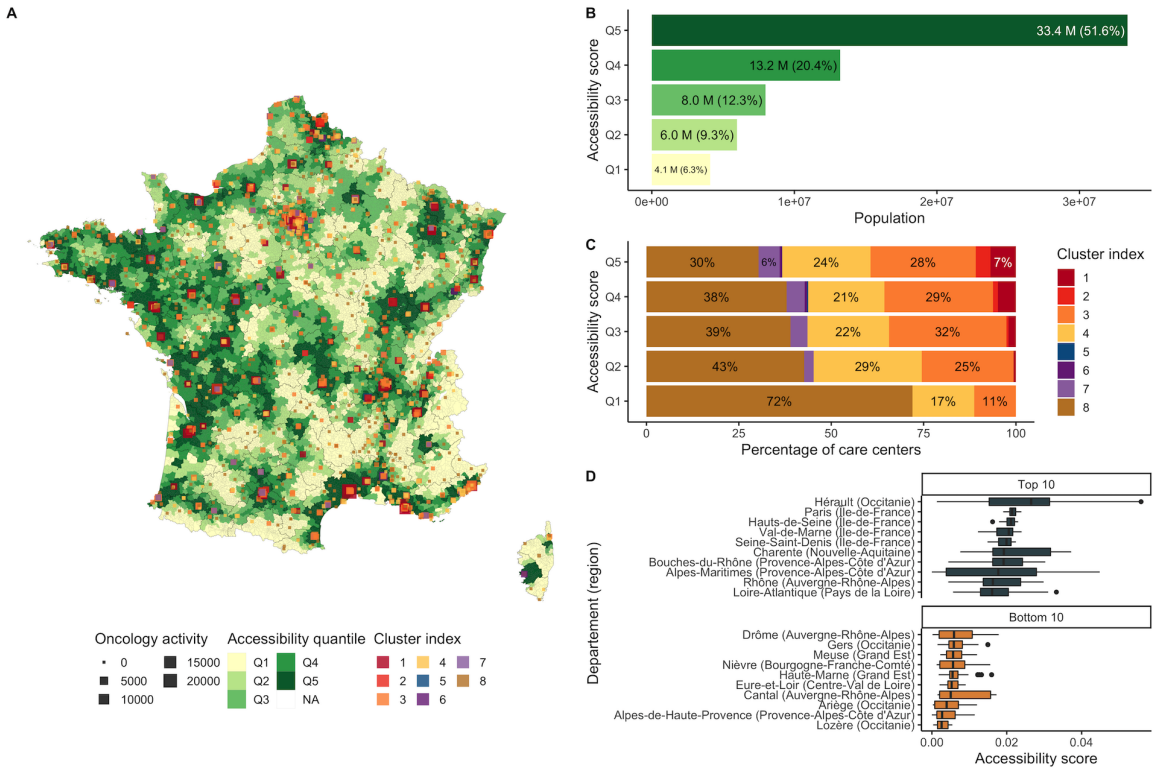


Figure 3.2: Distribution of the accessibility score computed with the E2SFCA, in metropolitan France. Plot (A) shows municipalities colored by accessibility quantile. The care centers are drawn as squares, colored by cluster, and sized by oncology activity. Plot (B) shows the total population by accessibility quantile. Plot (C) displays the percentage of care centers by cluster by accessibility quantile. Plot (D) shows the top 10 and bottom 10 list of the departments, ranked by median accessibility.

Accessibility score should be put into perspective with population density. Overall, the

denser municipalities have a median accessibility around 0.02. Municipalities with low population densities have more extreme values. Figure 3.3 compares accessibility and population density for three different regions: Provence-Alpes-Cote-d'Azur (A), Ile-de-France (B), and Bourgogne-Franche-Comté (C). Municipalities are displayed as squares, colored by accessibility quantile, and sized by population density. These regions show very different profiles. In Provence-Alpes-Cote-d'Azur (A), accessibility is essentially low in non-dense municipalities near the Alps. However, in Bourgogne-Franche-Comté (C), we see dense municipalities with poor accessibility scores, representing a large proportion of the region. We also drew similar maps (D, E and F) where municipalities are colored based on the average travel duration for patients with cancer in 2018. We see that the average travel time is higher in municipalities with poor accessibility scores.

Finally, we compared our accessibility score with the department exit ratio, by municipality. Department exit ratio is defined as the proportion of cancer patients who visited a care center outside from their department of residence and was computed using the PMSI database. In Provence-Alpes-Cote-d'Azur, the exit ratio is higher in departments with low accessibility scores and few oncology specialized care centers, as in Alpes-de-Haute-Provence and Hautes-Alpes. While the Var department has some oncology centers, exit ratio remains high since larger care centers are in Marseille and Nice.

Accessibility in Provence-Alpes-Cote-d'Azur region

We now focus on the region Provence-Alpes-Cote-d'Azur. This region is the far southeastern on the mainland. The region's population was 5,048 million in 2018. Its prefecture and largest city is Marseille. The region contains six departments. Bouches-du-Rhone, Var and Alpes-Maritimes are located on the coastline and gather the largest cities like Marseille, Nice, or Toulon. Alpes-de-Haute-Provence, Vaucluse, and Hautes-Alpes are inland departments, with a majority of rural and mountainous areas. Results are shown on Figure 3.4. By comparing maps (A) and (B), we confirm that the accessibility is maximum in denser areas of the region. Average patients travel time are displayed on map (C) and we drew the major

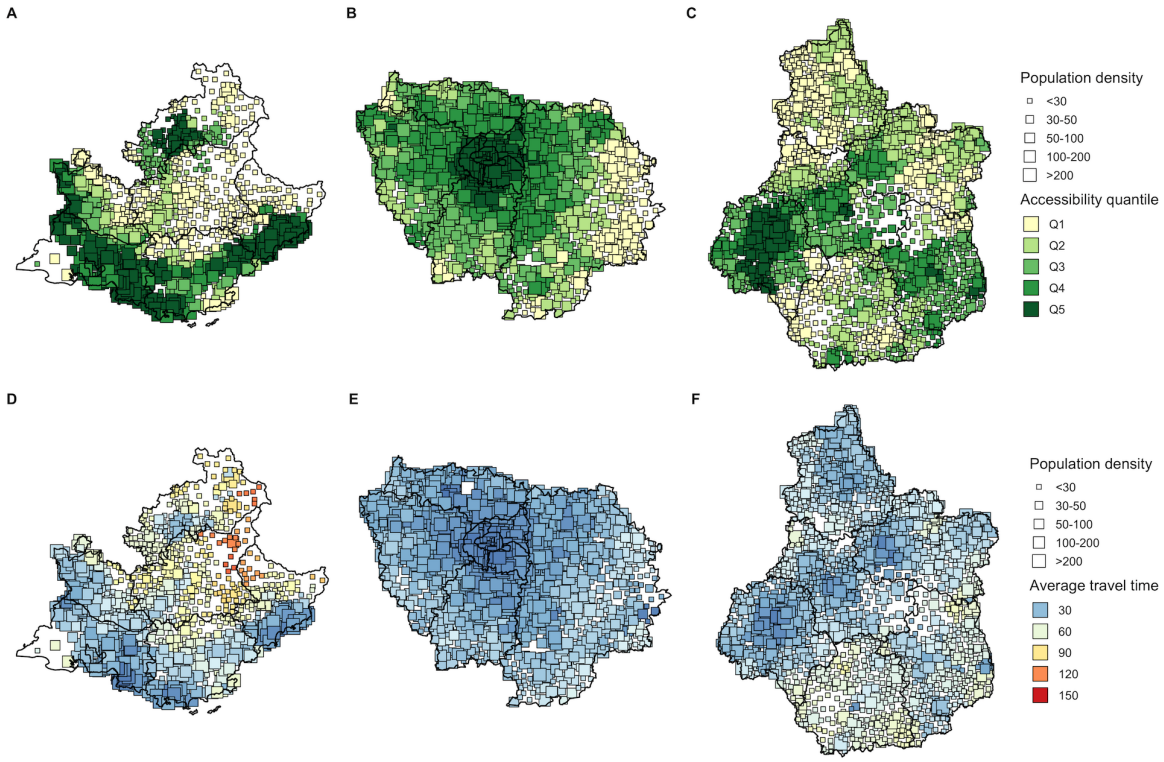


Figure 3.3: Comparison of population density with accessibility scores and patient average travel time for cancer pathways. Showing results in three regions: Provence-Alpes-Cote-d'Azur (A, D), Ile-de-France (B, E) and Bourgogne-Franche-Comté (C, F). Municipalities are drawn as squares, sized by population density and colored by either accessibility quantile (A, B, C) or patient average travel time (D, E, F).

roads (primary, motorway and truck) in red. The road system is well developed on the coast, rallying the larger cities of the region. However, driving from the rural areas in the Alps to the major cities is hard, resulting in higher travel times. The accessibility is unevenly spread within the departments, especially in Alpes-Maritimes where the distribution is multi-modal (D). There, cities like Nice and Cannes have large hospitals thus good accessibility, while the northern areas of the department are mostly mountains. Accessibility is higher in municipalities with dense populations, for all the departments (E). Finally, the average travel time decreases when the accessibility score increases. This makes sense since the accessibility score was computed based on the driving distance between population locations and care

centers. However, it confirms that patients living in poor accessibility zones effectively travel further to seek oncology care. In Bouches-du-Rhone, nearly all the municipalities have an average travel time lower than 30 minutes, while in Alpes-de-Haute-Provence, average travel times are rarely lower than 60 minutes (F).

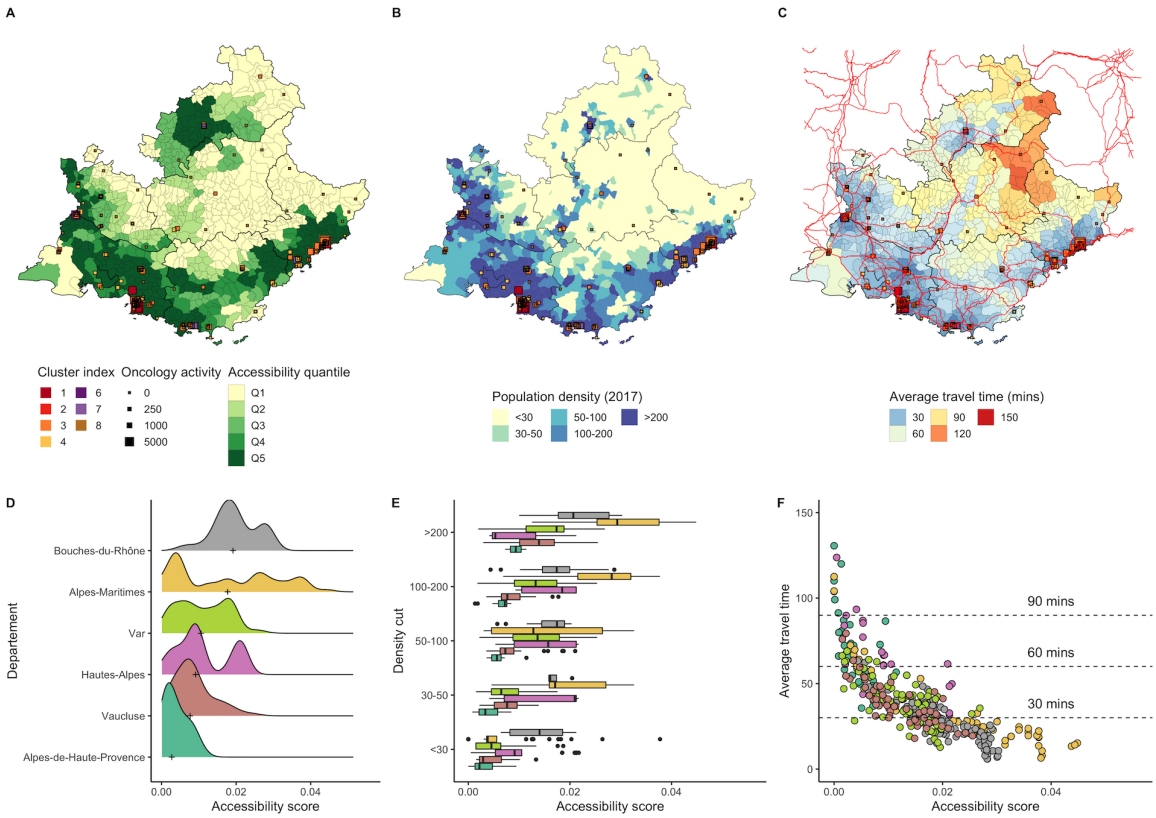


Figure 3.4: Accessibility distribution in Provence-Alpes-Cote-d'Azur region. Map (A) shows the region accessibility distribution per municipality. Map (B) displays the population density discretized in 5 bins. The map on plot (C) displays the average travel time for cancer pathways. Large roads (primary, motorway and trucks) are drawn in red. Plot (D) shows the accessibility distribution per department of the region. Plot (E) shows the accessibility distribution by municipality population density and department. Plot (F) compares the accessibility score from municipalities with the average travel time for cancer pathways.

Accessibility in Pays de la Loire region

The Pays de La Loire region is located in the west of France. It covers 32,082 km^2 which makes it the largest region in France, with a population of 3,806,461 (Insee) in 2019. In the region, one out of two inhabitants lives in rural areas, compared to one out of three on average in France. The Pays de la Loire is thus the 4th most rural region behind New Aquitaine, Brittany and Burgundy-Franche-Comté. The Pays de La Loire region is composed of 5 departments. The level of population living in rural communes varies according to the departments, but 4 departments out of the 5 are considered rural. In Vendée and Mayenne, two out of three inhabitants live in rural areas, in Maine-et-Loire 58% of the population resides in a rural commune and in Sarthe 56%. However, 29% of the region's population lives in a rural commune under the influence of a pole, compared to 20% in an independent rural commune. The city of Nantes, located in Loire-Atlantique in the east of the region, is the largest urban area in the region and has 303,382 inhabitants, as well as 961,521 inhabitants in its urban unit. The region has several cities with more than 100,000 inhabitants with Le Mans and its 143,325 inhabitants, Angers (151,520 inhabitants), followed by cities of about 50,000 inhabitants such as Saint-Nazaire, (68,200 inhabitants) Cholet, (54,200 inhabitants) and Laval (51,000 inhabitants). The Pays de la Loire has good accessibility with 51% of its population living in a territory with maximum accessibility and a low rate of its population living in territories with low or very low accessibility: 8.3% of its population resides in an accessibility score zone of Q2 and only 3.7% of its population in Q1. Thus, the maps show a good distribution of accessibility across the territory that varies proportionally with population density, with low accessibility areas corresponding to areas with low or very low population density. Travel time is also relatively evenly distributed across the region, with average travel times of 30 minutes, although depending on the department, a significant proportion of trips are between 30 and 60 minutes. A very small proportion of territories exceed 60 minutes of travel time. The territories with longer travel times are located in the Vendée department, mainly due to the coastal profile of the department and the islands that make it up, such as the Noirmoutier peninsula or the Ile d'Yeu, where travel times exceed 90 minutes and 120

minutes respectively.

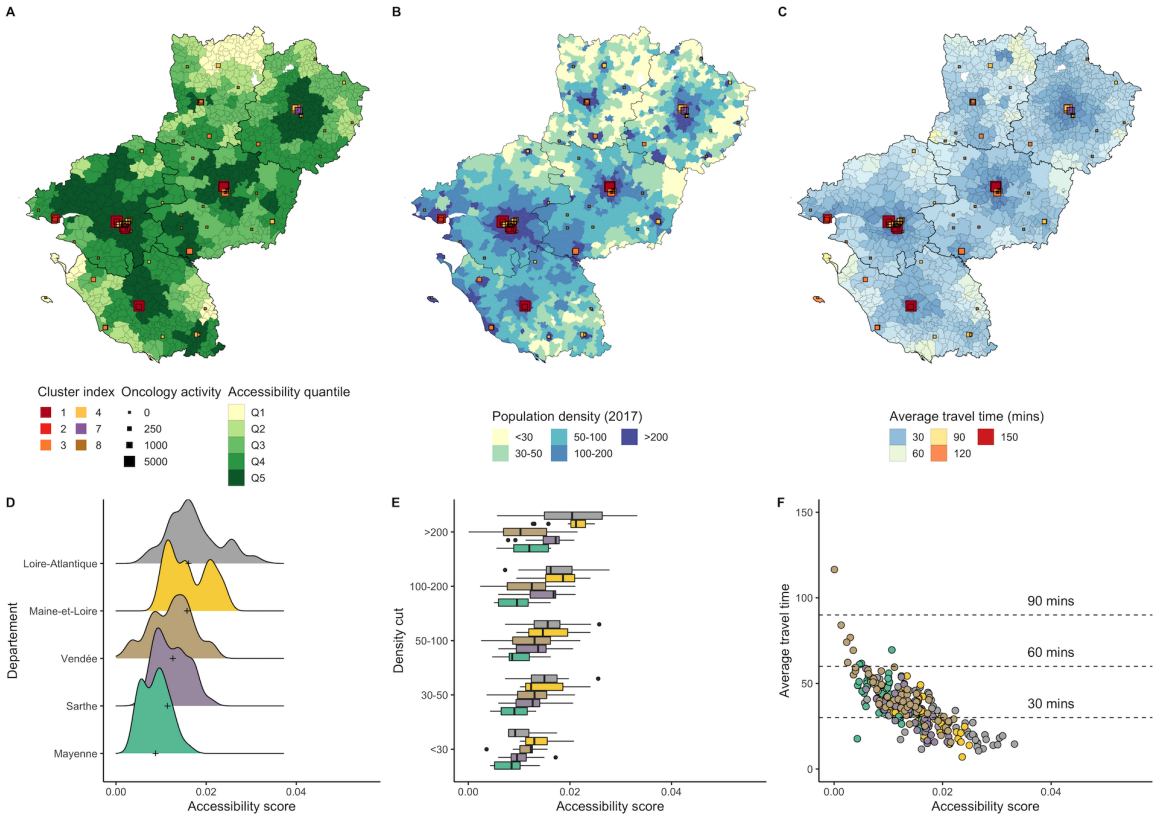


Figure 3.5: Accessibility distribution in Pays-de-la-Loire. The accessibility distribution in this region is high, and the amount of municipalities with Q5 accessibility score is very low. The median accessibility is the highest in Loire-Atlantique department, especially around Nantes; or in Maine-et-Loire near Angers. The lowest median accessibility is in Mayenne, where the main city is Laval. The accessibility is lower in the northern part of this department, where the population density decreases compared to the rest of the region.

Accessibility in Occitanie region

The Occitanie region is located in the South of France. It covers 72,724 kmš for a population of 5,933,185 (Insee) in 2019, for a population density of 81.6 inhabitants/kmš the 6th least dense region in metropolitan France. The rural territory in Occitanie represents 90% of the territory, mainly present in the mountain areas of the Massif Central and Pyrenees.

The urban space is mainly found along the coast and in the Garonne basin. 39% of the population lives in rural areas, i.e. 2.9 million inhabitants, and 9 of the 13 departments are considered rural. However, Occitanie is a largely urbanized territory with numerous urban centers in each department, the main metropolises being Toulouse and Montpellier. This region is the 5th most urbanized region of the metropolis and has more than fifty urban units of at least 10,000 inhabitants with several cities exceeding 70,000 inhabitants (Tarbes, Montauban, Albi). 4.4 million people live in the urban units, representing 76% of the population. Occitanie is composed of 13 departments. Three departments are among the most urbanized in the province and therefore have a strong demographic weight: Hérault (89% of the population residing in an urban unit), Pyrénées-Orientales (88%) and Haute-Garonne (87%). The Hérault department includes the city of Montpellier, but also Béziers, Sète and many small urban areas. The Haute Garonne includes the city of Toulouse, the fourth most populated commune in France (493,465 inhabitants) and with its rural areas are under strong pole influence. The Lot, Lozère and Gers are the least urbanized in France, with less than 40% of the population living in urban areas.

In this region, accessibility is not uniform across the territory. The areas with the highest accessibility scores are concentrated in the large urban areas and their catchment areas, notably in the center of the region around the city of Toulouse and Montauban in the Garonne basin, as well as along the coastline in the east of the region around the cities of Nîmes, Montpellier, Béziers, Narbonne and Perpignan. Also, if the most densely populated areas have a good level of accessibility, it can be seen that some medium-sized cities in the Occitanie region lack a good level of accessibility and even have low accessibility. This is particularly pronounced in the rural departments of the region (Lot, Gers and Lozère), as well as in Aude, Ariège and Hautes-Pyrénées. Indeed, many urban units have a low accessibility score (Q2) such as Auch (25,527 inhabitants) in the Gers, Foix (12,310 inhabitants) and Pamiers (29,340 inhabitants) in the Ariège, Rodez (47,868 inhabitants) in the Aveyron with a score of Q2/Q3, Cahors (24,279 inhabitants) in the Lot. Many areas of the region have long travel times of around 90 minutes if not 120 minutes on average. This is particularly true along the

border with Spain, which is characterized by its mountainous terrain. However, the Gers, Lot, Lozère, Aveyron and Hérault regions have average travel times of around 90 minutes. These high travel times are mainly associated with sparsely populated areas, although in the Hautes-Pyrénées department, average travel times of 90 minutes can be seen around the urban unit of Bagnères de Bigorre (13,213).

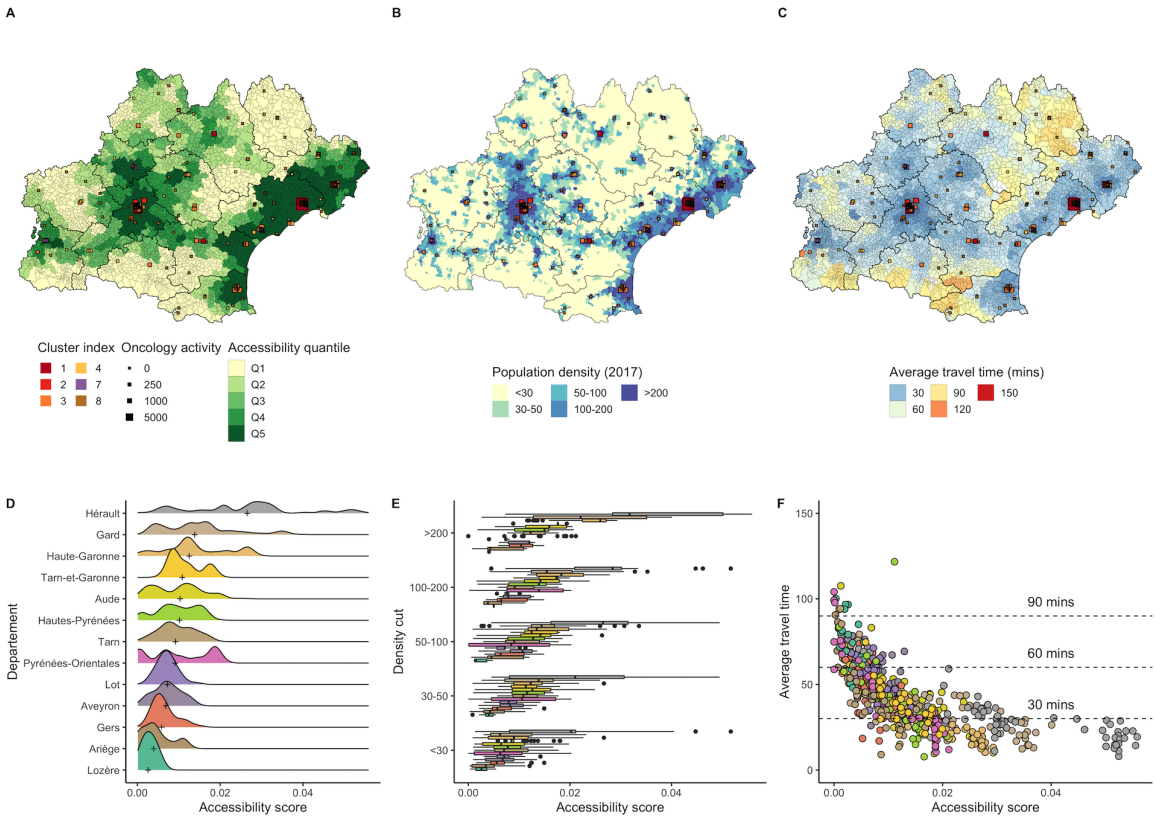


Figure 3.6: Accessibility distribution in Occitanie. The areas with the highest accessibility scores are concentrated in the large urban areas and their catchment areas, notably in the center of the region around the city of Toulouse and Montauban in the Garonne basin, as well as along the coastline in the east of the region around the cities of Nîmes, Montpellier, Béziers, Narbonne and Perpignan.

Accessibility in Nouvelle-Aquitaine region

The Nouvelle-Aquitaine region is located in the southwest of France. It covers an area of 84,036 km^2 which makes it the largest region in France, with a population of 6,010,289 (Insee) in 2019. The region is the third most rural region of France with half of its inhabitants living in a rural commune. The share of population in rural autonomous is significant compared to the national average but is similar to that of Brittany or Burgundy-Franche-Comté. Among the twelve departments of Nouvelle-Aquitaine , ten are predominantly rural, and two are predominantly urban: Gironde (71% of the population living in an urban commune) and Pyrénées-Atlantiques (62%). Nouvelle-Aquitaine is composed of 12 departments. The region's main metropolis, Bordeaux, with 260,958 inhabitants and 986,879 inhabitants in its urban unit, is located in the west of the region in the Gironde department. The region includes several intermediate cities with more than 70,000 inhabitants such as Limoges (130,876), Poitiers (89,212), Pau (75,627), La Rochelle (77,205), Mérignac (72,197), Pessac (65,245).

We notice accessibility disparities in this region. The areas with the highest accessibility scores are mainly located around the above-mentioned large and intermediate cities. Also, the areas with accessibility scores Q1 and Q2 are mainly located in territories with low or very low population density. Similarly, the Nouvelle-Aquitaine region seems to provide relatively widespread access to cancer care for its population. Indeed, 56% of its population is located in a zone with a maximum accessibility score of Q5, and 21.1% in a zone with a very good accessibility score of Q4. This leaves a smaller share of the population in areas of low accessibility (8.4% in Q2) and very low accessibility (6.3% in Q1). The average travel time is well distributed over the territory, with a majority of the territory covered by travel times between 30 and 60 minutes. It can be seen, however, that part of the territory has a good share of trips of less than 30 minutes (on average e 15 minutes) even in areas with average accessibility (score 0.2). A clear correlation can be seen between accessibility score and average travel time, with longer travel times in areas with low accessibility scores, but consequently less densely populated territories. The Landes and Lot-et-Garonne are the departments with the highest number of trips exceeding 60 minutes.

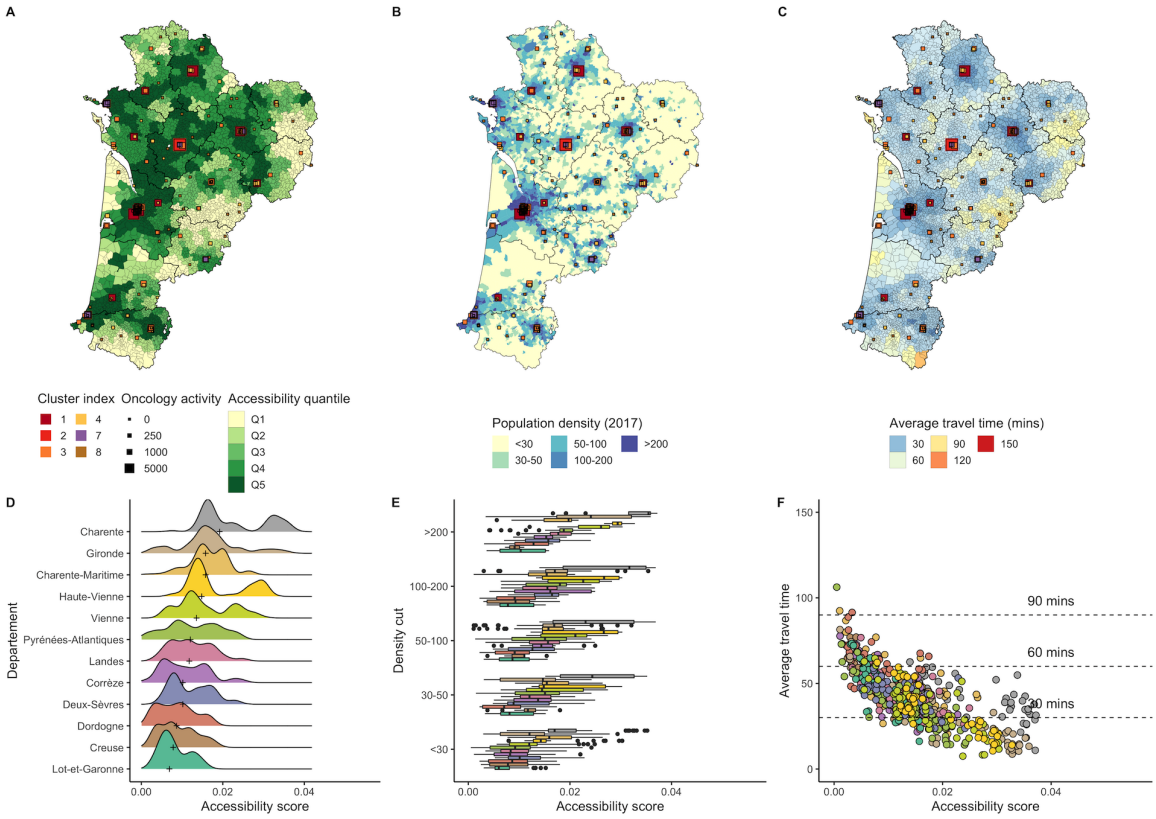


Figure 3.7: Accessibility distribution in Nouvelle-Aquitaine. The areas with the highest accessibility scores are mainly located around the above-mentioned large and intermediate cities. Also, the areas with accessibility scores Q1 and Q2 are mainly located in territories with low or very low population density. The Nouvelle-Aquitaine region seems to provide relatively widespread access to cancer care for its population.

Accessibility in Normandie region

The Normandie region is located in the north-east of France. It covers 29,905 kmš with a population of 3,325,032 inhabitants (Insee) in 2019. The Normandie region remains a fairly rural region with half of its inhabitants living in a rural commune (49% compared to 40% in the rest of France). The population living in a rural commune is clearly in the majority in Orne in the south of the region (73%), Manche in the west (68%) and Eure in the east (62%). However, more than half of the rural communes are not under the influence of a hub. Nor-

mandie is composed of 5 departments. The department of Seine-Maritime in the northeast of the region has two of the largest urban units in the region with more than 200,000 inhabitants: Rouen the most populous with 112,321 inhabitants and 471,893 in its urban unit as well as Le Havre with 172,366 inhabitants and 233,414 in its urban unit. The third urban unit of more than 200,000 inhabitants in the region is Caen with 206,973 inhabitants in its urban unit, located in Calvados. Normandie presents a rather average accessibility in terms of population density and accessibility ratio since 30.9% of its population lives in the best accessibility score almost equivalent to the percentage of population living in a territory with a Q3 score of 28.3%. Only 10.3% of its population lives in accessibility level Q1 and 9.2% in Q2.

We notice that the accessibility score is unevenly distributed. Although the areas with low or very low population density are the most affected by a low accessibility score of Q1 or Q2, we can still observe a fairly homogeneous distribution of the population on the territory, especially in the areas far from the urban units, and an accessibility that remains fairly low around Q2. The department of Calvados has the best distribution of accessibility over its entire surface. Whereas Orne, which is the most rural department in Normandie, has an accessibility score of Q1 except around the urban unit of Argentant. The same is true for the department of La Manche, which includes many areas of the territory with an accessibility score of Q1 or especially Q2 despite a higher population density, notably around the city of Cherbourg-Octeville and its surroundings with an accessibility score of Q3 or even Q2 for a city that nevertheless counts 35,545 and 81,423 in its urban unit (Figure 24). The average travel time is well distributed over the territory, with the majority of the territory covered by travel times of 30 minutes on average and below 60 minutes. It can be seen that the majority of trips in the departments of Seine-Maritime and Calvados are under 30 minutes, particularly in Calvados, unlike the department of Orne, the only department in the region whose trips are slightly over 60 minutes but still under 90 minutes.

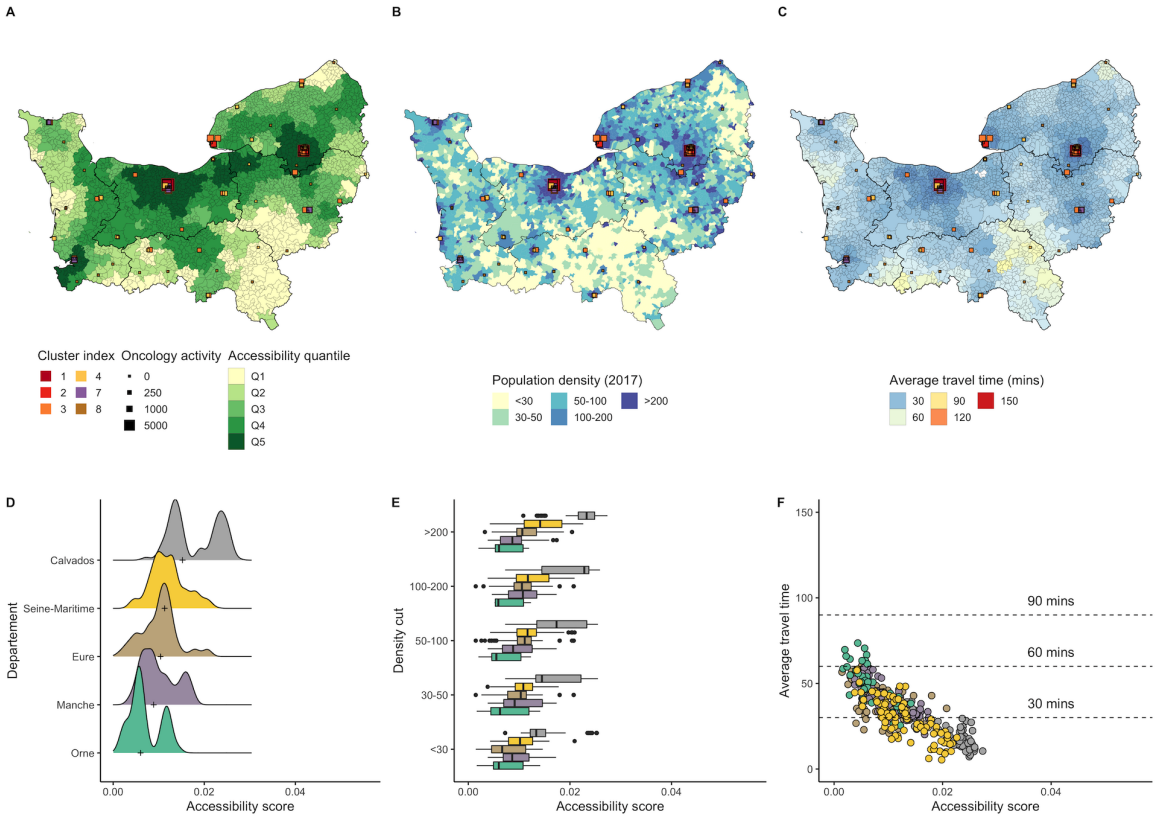


Figure 3.8: Accessibility distribution in Normandie.

Accessibility in Ile-de-France region

The Île-de-France (IdF) region is located in the center north of France. This one covers 12,012 km^2 for a population of 12,213,447 (Insee) in 2018. The IdF region is the most populated and dense of metropolitan France. Only 5% of the population lives in a rural commune, for the 671 rural communes cover 59% of the IdF territory. The majority of rural communes (85%) are under the influence of Paris. Île-de-France is composed of 8 departments, 4 departments in the inner suburbs and 4 departments in the outer suburbs. It is a special region because it includes the French capital, Paris, the leading French city in terms of demography and population density with 2,175,601 inhabitants in 2021. The city of Paris is also home to many specialized health establishments. The rural communes are far from the influence of Paris and are mainly located in the departments of the outer suburbs, three quarters of which

are in Seine-et-Marne. The most rural and least dense areas are therefore mainly located in the east of the region, particularly along the border to the east of the Seine-et-Marne department.

Île-de-France has good accessibility over the vast majority of its territory. Indeed, 63.8% of the population of IdF is located in an area with a maximum accessibility score, and almost no population is located in an area with a minimum accessibility score Q1 or even Q2. Also, although only 9% of the territory's surface is identified as having a Q5 score and 15% as having a Q1 score, the minimum accessibility zones are not very densely populated, which only affects a very small part of the region's population. Indeed, we observe that the only areas with a Q1 score are located in the eastern part of the region in the Seine-et-Marne department where the population density is very low. Moreover, travel time is uniform throughout the region with a very good level of travel time limited to an average of 30 minutes. The Ile-de-France region does not suffer from accessibility difficulties at any level for cancer treatments, regardless of location in the territory.

Accessibility in Hauts-de-France region

The Hauts-de-France region is located in the north of France. It covers 31,948kmš for a population of 6,005,000 (Insee) in 2019, or 9% of the metropolitan population. The region has retained a strong industrial footprint. It is the second most urbanized region after Ile de France with 89% of its population living in a large urban area. However, 83% of the region's municipalities are considered rural (including autonomous rurality and rurality under the influence of a pole in a peri-urban area), with 29% of the region's population living in a so-called rural municipality. The Hauts-de-France is composed of 5 departments. In the department of Nord in the north of the region, particularly urbanized and densified, is the city of Lille which has 1,411,571 inhabitants in its metropolis. Amiens in the department of Somme is the second most populated urban area in the region.

The accessibility zones are relatively evenly distributed over the territory, although the best accessibility in this department is mainly in the urban and peri-urban area of Lille. Travel

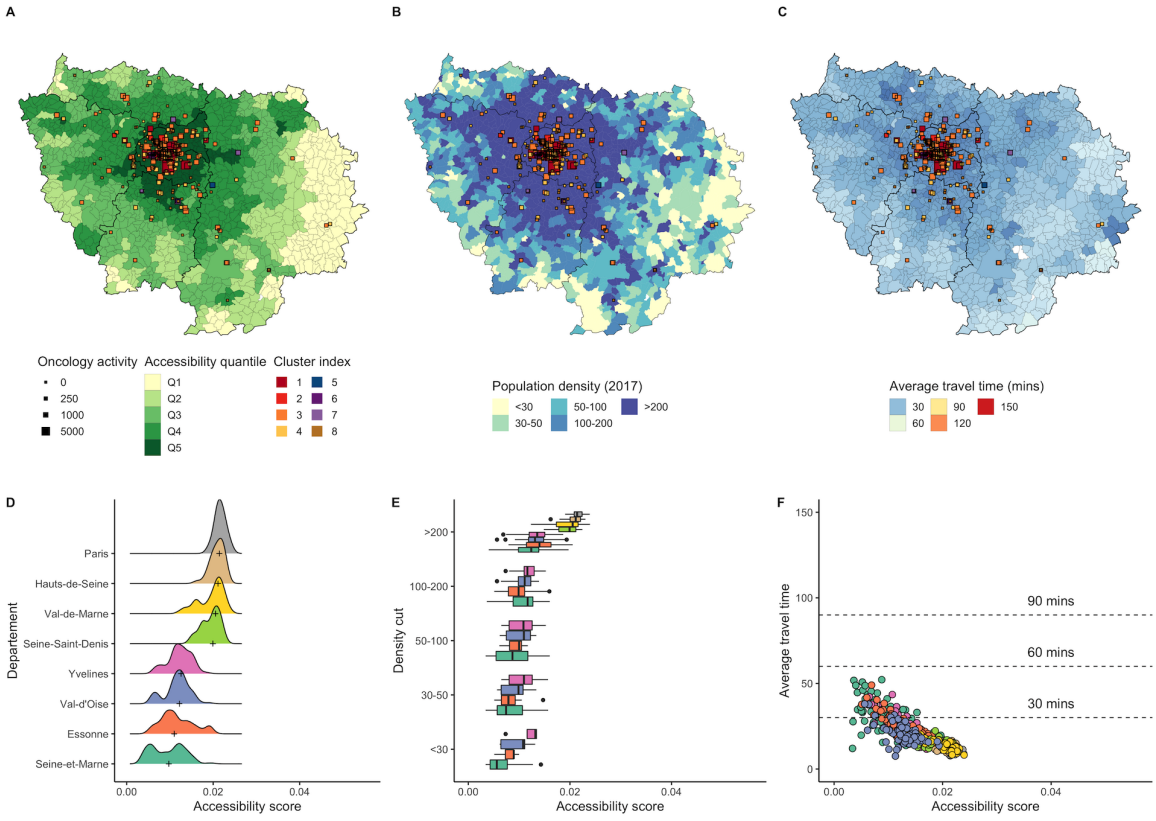


Figure 3.9: Accessibility distribution in Ile-de-France.Île-de-France has good accessibility over the vast majority of its territory. Indeed, 63.8% of the population of IdF is located in an area with a maximum accessibility score, and almost no population is located in an area with a minimum accessibility score Q1 or even Q2. Also, although only 9% of the territory's surface is identified as having a Q5 score and 15% as having a Q1 score, the minimum accessibility zones are not very densely populated, which only affects a very small part of the region's population.

time averages 30 minutes over most of the region, with the exception of the northern end of the region in the Aisne department and the northeastern part of the same department, where travel time averages 60 to 90 minutes. The population density is also low in these areas, the Aisne being the least populated department in the Hauts-de-France region. Only 4.4% of the population of Hauts-de-France is located in an area with an accessibility quantile of Q1 and 16.6% in Q2. It is possible to perceive that certain parts of the territory with a

medium (100-200) to high (>200) population density have an accessibility qualified at Q2 and Q3, which implies a difficulty of accessibility of optimal care for certain segments of the population. In fact, despite a low population rate in Q1, only 32.5% of the regional population is located in an area with the highest quantile of accessibility Q5. These observations allow us to consider that the improvement of accessibility to optimal care in this territory could be easily optimized because the initial care offer is already well distributed in the territory with accessibility zones Q4 and Q3, which together account for 46.4% of the population.

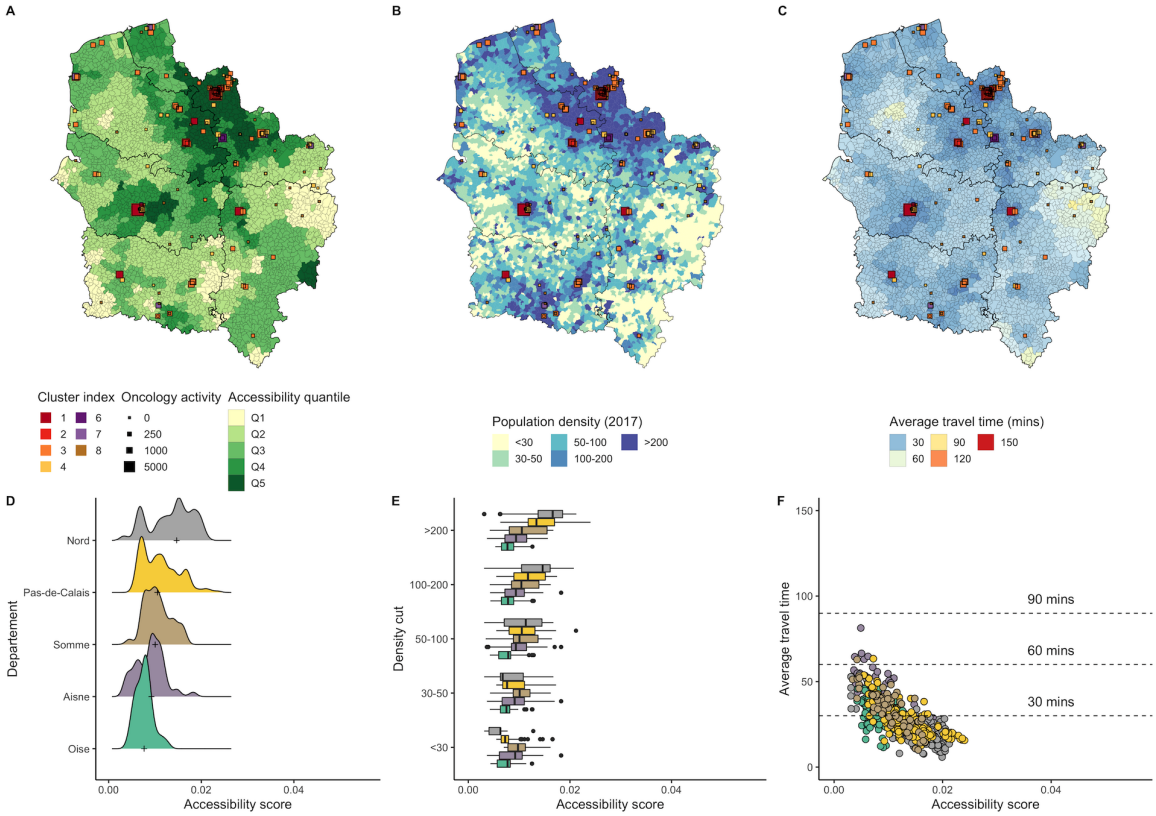


Figure 3.10: Accessibility distribution in Hauts-de-France. The accessibility zones are relatively evenly distributed over the territory, although the best accessibility in this department is mainly in the urban and periurban area of Lille. Travel time averages 30 minutes over most of the region, with the exception of the northern end of the region in the Aisne department and the northeastern part of the same department, where travel time averages 60 to 90 minutes

Accessibility in Grand Est region

The Grand Est region is located in the east of France. It covers 57,433 km^2 for a population of 5,556,219 (Insee) in 2019. 39% of the population resides in a rural commune (i.e., a commune with low or very low density). 61% of the population resides in urban areas, 22.8% in peri-urban rural areas and 16.2% in autonomous rural areas, moreover nearly 80% of the regional surface is dedicated to agriculture and forestry. The Grand Est is composed of 8 departments. The departments of Meuse and Haute-Marne central to the region are among the most rural departments in France with respectively 74% and 67% of their population living in rural areas (peri-urban and autonomous), while the departments of Haut-Rhin, Bas-Rhin, Meurthe-et-Moselle and Moselle have more than 60% of their population living in urban areas (2018, Insee). The department of Marne in the west of the region is home to Reims, the most densely populated city in the region after Strasbourg.

The accessibility is high in the eastern half of the region in the departments of Moselle, Meurthe et Moselle, Bas-Rhin, Haut-Rhin, particularly around the large agglomerations (Strasbourg, Nancy, Metz, Colmar). Indeed, 41% of the population of the Grand Est is in an accessibility zone of Q5 and only 7.5% in a Q1 zone. The lack of accessibility in the western part of the region is more pronounced due to the low or very low density areas that are more common in these departments. Also, the link between population density and accessibility is visible and reinforced by the consideration of average travel times. Travel times are almost uniformly distributed over the entire territory, with little or no travel time exceeding 30 minutes; travel times of 60 minutes on average are limited and those of 90 minutes are very limited. These times are most prevalent in the western half of the region in the very low density areas but mostly in the less demographically dense areas. The poor accessibility for the city of Charles-Ville-Mézière (46,436 inhabitants in 2019) is more worrying in view of its demographic density. However, it can be observed that the coverage of maximum accessibility for the majority of the population does not necessarily require a spatial accessibility spread over the surface of the region, since the Grand Est has only 13.5% of the surface of its territory considered as Q5 accessibility, but covers the needs of maximum accessibility

for 41 of its population. Thus, the urban nature of the population of the Grand Est seems to be a determining factor in maximizing accessibility to care.

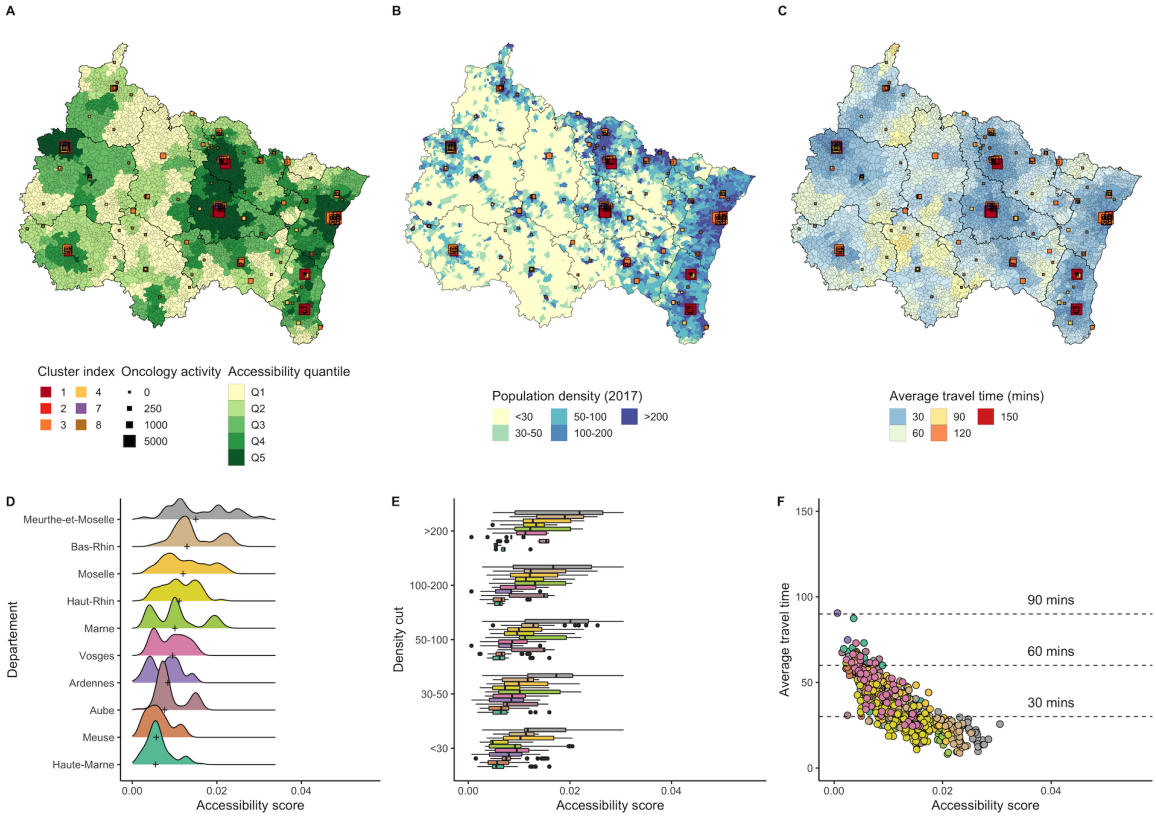


Figure 3.11: Accessibility distribution in Grand-Est. We notice good accessibility scores in the eastern half of the region in the departments of Moselle, Meurthe et Moselle, Bas-Rhin, Haut-Rhin, particularly around the large agglomerations (Strasbourg, Nancy, Metz, Colmar). Indeed, 41% of the population of the Grand Est is in an accessibility zone of Q5 and only 7.5% in a Q1 zone. The lack of accessibility in the western part of the region is more pronounced due to the low or very low density areas that are more common in these departments.

Accessibility in Centre-Val de Loire region

The Centre-Val-de-Loire region is located in the center west of France. It covers an area of 39 151 km^2 with a population of 2 573 180 (Insee) in 2019. The region is one of the most rural regions of France, with 90% of its territory occupied by rural municipalities and 1 in 2

inhabitants living in a rural municipality (49%). 27% of the population (700,000 inhabitants) live in a rural commune under the influence of a major pole and nearly 22% (of 570,000) outside the area of attraction of such a pole. However, the CVdL includes two metropolitan areas, Orléans in the department of Loiret and Tour in Indre-et-Loire, which together account for one-third of the regional population. Paris also has an influence on the region, affecting 184,000 inhabitants under its influence, i.e., 7% of the CVdL population. Thus, the majority of the population (90%) lives in an attractive urban area. The Hauts-de-France is made up of 6 departments. The department of Indre-et-Loire includes and Loiret includes the two metropolitan areas of the region Tour with 137,665 inhabitants and Orleans with 288,229 inhabitants in 2019.

The accessibility of the whole region is relatively lower than in other regions observed so far. Many areas have a low or very low accessibility score despite a medium population density. Areas with low or very low population density can have a very low accessibility score, although low-density areas of the Cher have a score around the Q3 quantile. Only the city of Tour and its vicinity shows a maximum level of accessibility, as well as some surrounding parcel areas in the department of Loir-et-Cher around the city of Blois and in the department of Cher around Bourges. Even the city of Orleans has an accessibility score of Q4 despite the presence of level 1 clusters. The CVdL has the particularity of being the only French region without a CLCC on its territory. The closest CLCC are those in adjacent regions, in Paris in the Île-de-France and Anger in Normandy. We can deduce that in order to access a specialized center, the inhabitants of this region have to leave the region. We can see that the level 1 clusters in the region are located in Tour, Orléans and Chartes. The departments in the south of the region have lower level clusters, with the Cher having only a level 3 and a level 7 cluster. This is reflected in the travel times which are rather homogeneous and low in the northern and central departments with average travel times of 30 minutes, while the southern departments, Indre and Cher have much higher travel times throughout their territory, around 60 minutes and 90 minutes.

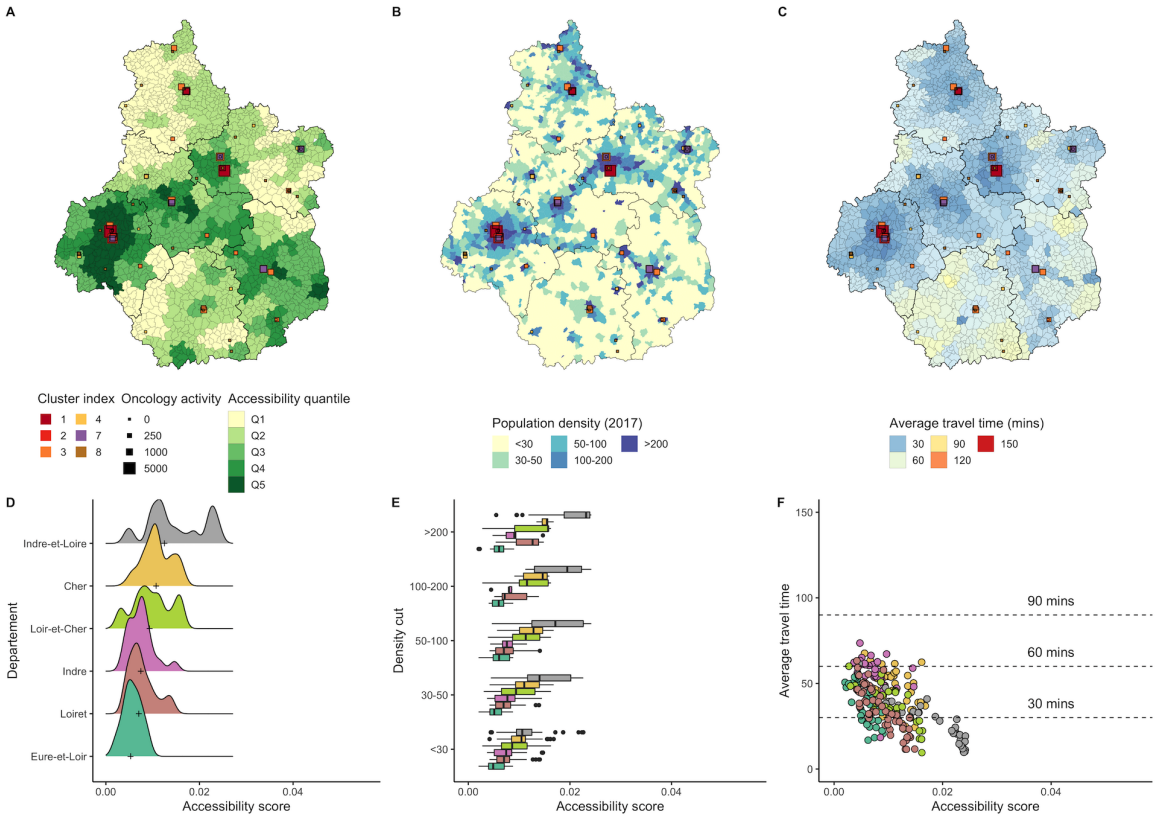


Figure 3.12: Accessibility distribution in Centre Val de Loire. The accessibility of the whole region is relatively lower than in other regions observed so far. Many areas have a low or very low accessibility score despite a medium population density. Areas with low or very low population density can have a very low accessibility score, although low-density areas of the Cher have a score around the Q3 quantile. Only the city of Tour and its vicinity shows a maximum level of accessibility, as well as some surrounding parcel areas in the department of Loir-et-Cher around the city of Blois and in the department of Cher around Bourges.

Accessibility in Bretagne region

The region of Bretagne is located in the west of France, on the Atlantic coast. It covers 27,208 kmš, making it the largest region in France, with a population of 3,354,854 (Insee) in 2019. More than half of the Breton population (53.7%) resides in a rural commune, so Bretagne

is the second most rural region of metropolitan France after Burgundy-Franche-Comté. The Breton rural area is characterized by longer travel times to everyday services. 25.7% of the inhabitants of very sparsely populated autonomous areas have to travel more than 10 minutes on average to access them, and for 68.6% of them, the average journey takes between 7 and 10 minutes. However, a major part of the population lives in an attractive urban area, i.e. 87% of the region's population. Bretagne is composed of 5 departments. The main metropolis of the region is Rennes with 215,366 inhabitants and 364,133 inhabitants in its urban unit, the first agglomeration of the department of Ille-et-Vilaine, followed by Brest which is located in the department of Finistère with 139,926 inhabitants.

Bretagne has very good accessibility with 57.7% of its population living in a territory with maximum accessibility and above all a very low rate of its population in territories with low or very low accessibility with 5.1% of its population in Q2 and only 1.5% of its population in Q1. Also, the maps show a good distribution of accessibility throughout the territory, with variations often related to the territory's population density ratio. Travel times reflect the level of accessibility, with many travel times less than 30 minutes and some travel times between 30 and 60 minutes but very rarely more. However, Morbihan has a relatively high proportion of trips between 30 and 60 minutes, including rare areas where travel times exceed 90 minutes, particularly due to the department's profile, which includes certain islands such as Belle-Île, which have travel times of over 120 minutes.

Accessibility in Bourgogne-Franche-Comte region

The Bourgogne-Franche-Comté (BFC) region is located in the center-east of France. It covers 47,784 km^2 for a population of 2,805,580 (Insee) in 2019 with 1,242,882 active people. In 2018 the BFC is considered the first rural region of France with more than half of its population (1.5 million people) residing in rural areas. The BFC is composed of 8 departments. The departments of Yvonne, Nièvre to the west, Saône-et-Loire and Jura to the south, have a particularly rural and agricultural landscape without dense urban areas, especially for Saône-et-Loire. In the department of Côte-d'Or is located Dijon, the largest and most densely pop-

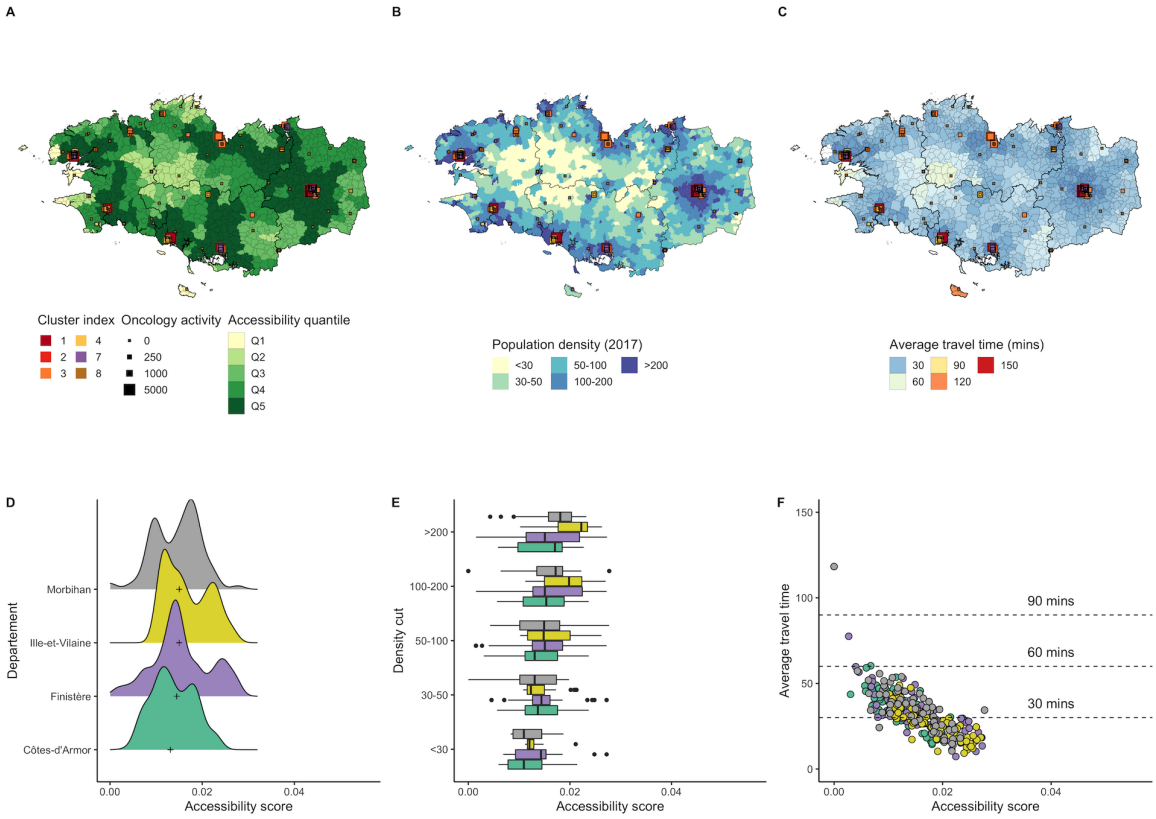


Figure 3.13: Accessibility distribution in Bretagne.

ulated city in the region with 158,002 inhabitants, ahead of the city of Besançon with its 117,912 inhabitants, which is located to the east in the department of Doub. In total, the BFC region has 3,704 municipalities, 26 of which have more than 10,000 inhabitants.

The departments of Côte-d'Or and Doubs have the best accessibility, especially around densely populated urban areas such as Dijon or Besançon. Some areas of the region have a low accessibility quantile Q1 and Q2 which cover 37.3% and 16.4% respectively of the regional territory, i.e. more than half (53.7%) of the area is recognized with a level of accessibility to cancer care. The areas with low or very low accessibility are located mainly in rural areas and with low or very low population density, except for the eastern border of the Doubs, which has more densely populated areas, but with more mountainous terrain, with a quantile 1 accessibility. In each department, accessibility is best in the urban areas and their surroundings. The travel time shows an unequal distribution of access to health care in

the territory, since the majority of municipalities have an average travel time of 30 minutes, but large areas of the region show average travel times of 60 or 90 minutes, particularly in the departments of Nièvre and Yvonne, although in the less densely populated areas of these departments, or travel times exceeding 120 minutes in the east of the Jura at the Swiss border, which is, however, likely to be a more mountainous terrain

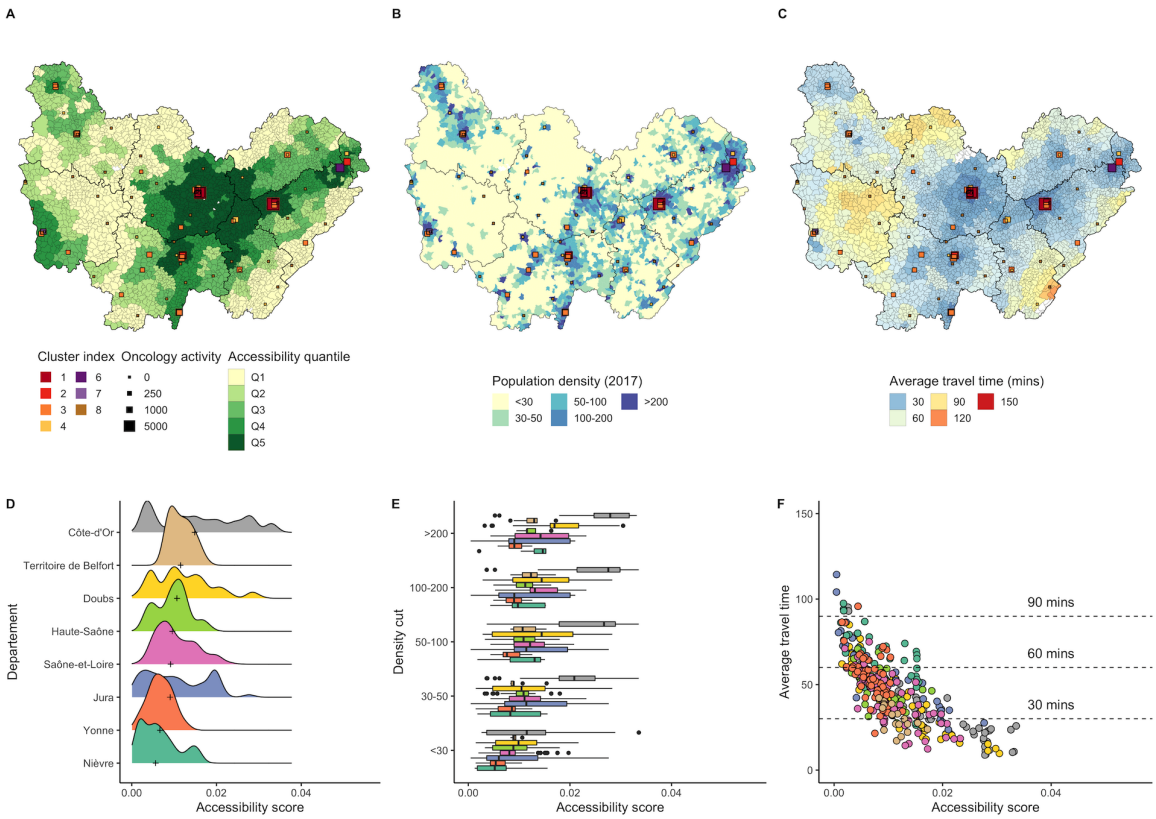


Figure 3.14: Accessibility distribution in Bourgogne-Franche-Comté. The departments of Côte-d'Or and Doubs have the best accessibility, especially around densely populated urban areas such as Dijon or Besançon. Some areas of the region have a low accessibility quantile Q1 and Q2 which cover 37.3% and 16.4% respectively of the regional territory, i.e. more than half (53.7%) of the area is recognized with a level of accessibility to cancer care. The areas with low or very low accessibility are located mainly in rural areas and with low or very low population density, except for the eastern border of the Doubs, which has more densely populated areas, but with more mountainous terrain, with a quantile 1 accessibility.

Accessibility in Auvergne-Rhone-Alpes region

The Auvergne-Rhône-Alpes (ARA) region is located in eastern France. It covers 69,711 km^2 for a population of 7,994,459 (Insee) in 2018, representing 12.3% of the metropolitan population, i.e. the most populated region in France. The ARA is the main mountain region of France with 2.2 million people residing in a municipality classified as a mountain area, with more than half in the regional part of the Massif Central which is distributed in a diagonal of low population density, while the population of the Alpine massif is concentrated in the urbanized and more densely populated parts at the bottom of valleys. In the ARA, 35% of the population lives in a rural commune, the provincial metropolitan average being 33%, and these communes cover 89% of the region's surface area. The ARA is composed of 12 departments. The Rhône department in the northern center of the region includes the city of Lyon, the second largest city in France, which has 1,411,571 inhabitants in its metropolis. The eastern departments, Savoie, Haute-Savoie, Isère, Drôme, constitute the mountainous areas of the region. Of the twelve departments, five are considered 'essentially rural': Cantal (74% of the inhabitants live in rural communes), Haute-Loire (70%), Ardèche (60%), Allier (58%) and Ain (50%).

If we look at the maps, we can see that the areas with the lowest accessibility are mainly located in areas with low or very low density, particularly along the mountainous border in the east of the region in the departments of Haute-Savoie, Savoie, Isère and Drôme. It is possible to observe a good distribution of accessibility in the central, northern and north-western part of the region, particularly around the large agglomerations such as the city of Lyon, Clermont-Ferrand, Moulins, Grenoble and Aurillac. The three southern departments, Haute-Loire, Ardèche and Drôme, are less accessible than the other departments in the region. Above all, it can be observed that the mountainous terrain tends to have a strong impact on accessibility to care, since travel times in these areas, particularly for the departments of Drôme and Savoie, reach an average of 120 minutes if not 150 minutes. In the mountainous departments of the east, the valleys that contain the urban centers with the highest population density, such as Chambéry, Grenoble and Annecy, are the most favor-

able accessibility centers in these departments. Despite its mountainous nature, 51.1% of the Auvergne-Rhône-Alpes region is located in an accessibility zone Q5 compared to 8% in an accessibility zone Q1.

3.3 Conclusion

In this section, we described our method to compute the oncology accessibility score given to every municipality in metropolitan France. This score was obtained by using the Enhanced Two Step Floating Catchment Area (E2SFCA) algorithm, with oncology activity as supply, municipality population as demand, and driving car duration as impedance metric. Specific attention should be given to municipalities with very poor access to oncology care centers. While we saw that most of the population lives in high accessibility areas, around 6% of the population lives in the bottom 20% accessibility quantile. Among these municipalities, some are very rural and mountainous like those in the Alpes-de-Haute-Provence in Provence-Alpes-Cote-d'Azur region. Such areas cannot be expected to have a very good healthcare coverage. By contrast, the case of suburban areas with relatively dense population and poor accessibility should be addressed more easily. Our optimization algorithm can help driving public health policies, as it effectively identifies areas where accessibility could grow, by allocating additional oncology activity to a restricted number of care centers. The proposed growth factors are indicative and do not have to be effective within a year, as it represents a considerable effort for care centers to increase their activity. Our oncology accessibility score is deliberately non-specific to cancer type. This score is meant to outline how easy it would be for a population location to reach a first entry point for oncology care. Here, we are only focusing on surgery, chemotherapy, and radiotherapy treatments. The same technique could be used on a specific cancer type, the method will remain the same, only the supply variable used in the accessibility score will change. We should mention that SA is better suited for pathologies that are relatively well handled across the whole country. Accessibility for rare diseases like pediatric cancer or complex cancers that re-quire a

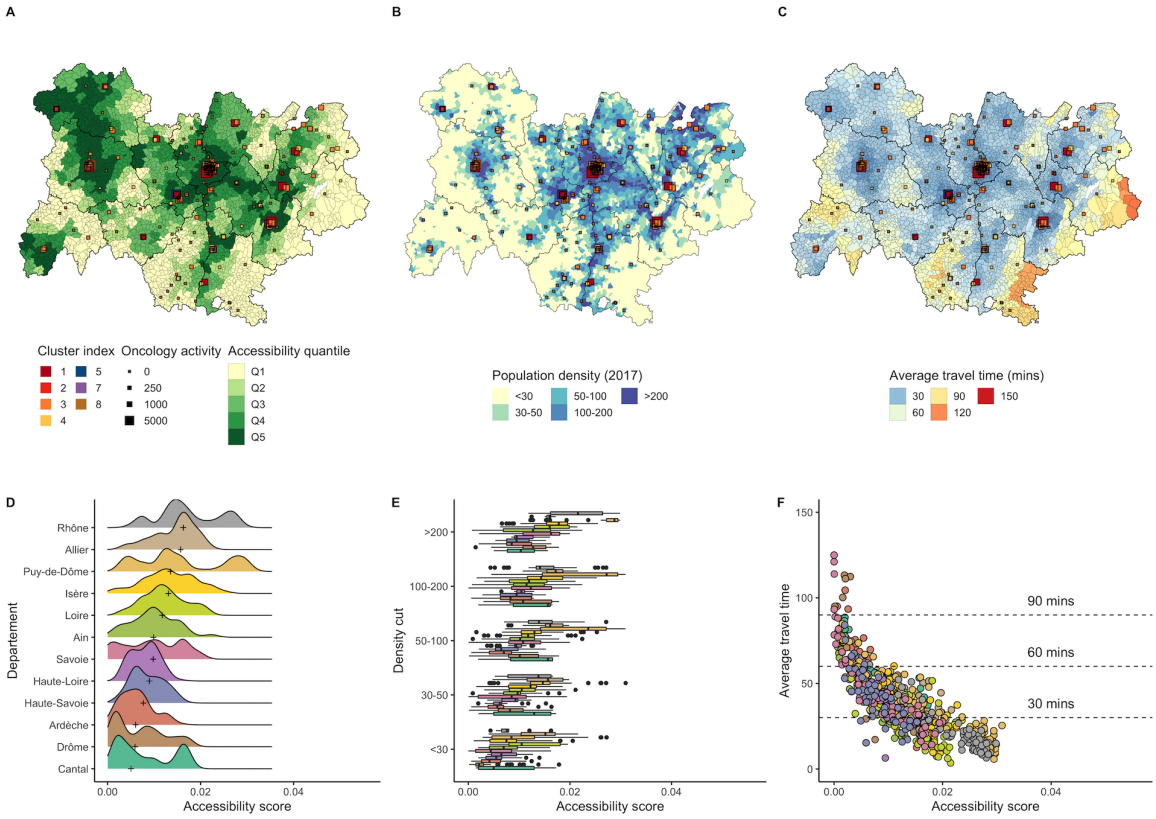


Figure 3.15: Accessibility distribution in Auvergne-Rhone-Alpes. The areas with the lowest accessibility are mainly located in areas with low or very low density, particularly along the mountainous border in the east of the region in the departments of Haute-Savoie, Savoie, Isère and Drôme. It is possible to observe a good distribution of accessibility in the central, northern and north-western part of the region, particularly around the large agglomerations such as the city of Lyon, Clermont-Ferrand, Moulins, Grenoble and Aurillac. The three southern departments, Haute-Loire, Ardèche and Drôme, are less accessible than the other departments in the region. Above all, it can be observed that the mountainous terrain tends to have a strong impact on accessibility to care, since travel times in these areas, particularly for the departments of Drôme and Savoie, reach an average of 120 minutes if not 150 minutes.

specific expertise is less informative because only a handful of care centers are indicated. Similarly, we could compute an accessibility score that is focused on specific kinds of stays:

our web application lets the user pick between surgery, chemotherapy, or radiotherapy as supply variable.

Chapter 4

Catchment Area MaximizatION (CAMION)

This chapter is part of a research article currently under submission. The preprint is available on medrXiv.

4.1 Methods

4.1.1 Overall optimization

We model the problem as an optimization task. In our case, we want our optimization algorithm to find new care centers capacities given some constraints, so that the total accessibility is maximum. We apply optimization on a given region only, rather than on the whole metropolitan France. We chose this approach because healthcare planning is handled regionally rather than nationally. We show below that our optimization problem is a Linear Programming (LP) problem. In its standard form, LP finds a vector x that maximizes $c^T x$ under constraints $Ax \leq b$, where A is a matrix and b a vector. Boundaries can be set to x such as $x \geq 0$. Consider x_u the new capacity of a care center u, to be computed by the algorithm. Let Q_u and W_u be two vectors of size m, defined as follows:

$$Q_u = \sum_{s=1}^{r} W_s \sum_{i, d_{iu} \in I_s} P_i \tag{4.1}$$

$$W_u = \sum_{s=1}^{r} \sum_{i, d_{iu} \in I_s} W_s \tag{4.2}$$

Intuitively, Q_u is the weighted population that has access to the care center u, and W_u is the sum of weights of municipalities that have access to u. We can compute the total accessibility as a sum on the m care centers:

$$\sum_i A_i = \sum_i \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u}{Q_u}$$

$$\sum_i A_i = \sum_i \sum_{i, d_{iu} \in I_s} W_s \frac{S_u}{Q_u}$$

$$\sum_i A_i = \sum_u \frac{S_u}{Q_u} \sum_s \sum_{i, d_{iu}} W_s$$

$$\sum_i A_i = \sum_u \frac{S_u}{Q_u} W_u \tag{4.3}$$

Equation (4.3) can be rewritten in the LP standard form with:

$$c = \frac{W_u}{Q_u}$$

$$x_u = S_u$$

$$b \geq \sum_u x_u$$

$$x_{u_{\text{min}}} \leq x_u \leq x_{u_{\text{max}}}$$

The user-defined parameters are b, $x_{u_{\text{min}}}$ and $x_{u_{\text{max}}}$. b is the total capacity to be shared across all the care centers. $x_{u_{\text{min}}}$ and $x_{u_{\text{max}}}$ are the capacity boundaries for care center u. If b is set to the current total capacity, a care center cant be grown unless another one is decreased. If $b > \sum_u x_u$, the capacity of care centers can be increased without decreasing other centers. We know how to solve LP and we used the SciPy [166] implementation of the revised simplex method as explained in [167]. We now detail how we set the user-defined parameters to apply the LP algorithm to our specific case. The additional capacity was set as +3% of the overall activity of the region's care centers: $b = 1.03 \times \sum_u x_u$. The choice of the boundaries $x_{u_{\text{min}}}$ and $x_{u_{\text{max}}}$ is crucial and must be realistic. We studied the hospitals activity on the past four years (2016 to 2019) to retrieve the average growth percentage of a care center. The growth percentage is computed as follows: $(S_{2019} - S_{2016})/S_{2016}$. Among the care centers that grew and who had an existing oncology activity, the mean growth percentage was 23%. Hence, we set $x_{u_{\text{max}}}$ as +20% of the care center capacity. Regarding $x_{u_{\text{min}}}$, we set the boundary based on the cluster of the care center. For the three most specialized clusters, we set their $x_{u_{\text{min}}}$ equal to their current activity. We did this to prevent the algorithm from decreasing the most specialized and well-equipped care centers. Regarding the care centers from the other clusters, $x_{u_{\text{min}}}$, so that they could be emptied if need be. Finally, we set $x_{u_{\text{max}}}$ if the care center belongs to the least specialized cluster. The new capacities are indicative and should be further investigated to make sure they are relevant. Especially when setting an existing oncology activity to 0.

4.1.2 Maxi-min optimization

We now want to maximize the minimum accessibility, meaning that the facilities capacities will be increased to develop the areas where the accessibility is minimum in priority. Let z be the minimum accessibility score.

$$z = \min_{i=1,...,n} A_i \tag{4.4}$$

$$z \leq A_i \text{ for all } i = 1, ..., n \tag{4.5}$$

Let x_u be the capacity increase for facility u, whose current capacity was S_u. A facility with an unchanged capacity will have $x_u = 0$. The accessibility score A_i at municipality i computed with the E2SFCA algorithm can be written as:

$$A_i = \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u + x_u}{Q_u}$$

Replacing A_i with this previous formulation in Equation (4.5) brings the following:

For all $i = 1, ..., n$:

$$z \leq \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u + x_u}{Q_u}$$

$$z - \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{x_u}{Q_u} \leq \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u}{Q_u}$$

We can add these new n equations as constraints to the optimization problem, as well as the other constraints. The Linear Programming problem is now framed as the maximization of $c^T x$ with $c = (1, ..., 0)$ and $x = (z, ..., 0)$, both of size $m + 1$ with m the number of facilities. The constraints are:

For all $i = 1, ..., n$:

$$z - \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{x_u}{Q_u} \leq \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u}{Q_u}$$

$$\sum_{u} x_u \leq b \text{ for a given budget } b$$

$$x_{u_{\text{min}}} \leq x_u \leq x_{u_{\text{max}}}$$

4.2 Results

We now present the outcomes of our optimization algorithm, on every region in metropolitan France. We chose to run the overall optimization approach, because it led to better results. Indeed, with the maxi-min approach, the municipalities with low population densities and few hospitals were targeted first. Since these municipalities often have access to non specialized hospitals, the only lever we had was to develop these smaller hospitals, which could be very costly. The algorithm was ran on every region and the additional number of stays was set to 3% of the current region's overall activity and capped care centers to a 20% maximum growth.

4.2.1 Optimization results in metropolitan France regions

Provence Alpes Cote d'Azur

We allocated 3,221 new stays in this region, corresponding to 3% of the overall activity. The median accessibility in the region went from 0.0093 to 0.0103, a 11.1% increase. The results are shown on Figure 4.1. Map (A) displays the accessibility delta ($A_{i_{\text{after}}} - A_{i_{\text{before}}}$) as well as the care centers eligible to grow. Centers from cluster 8 were hidden since we considered that they couldn't provide any oncology activity. The algorithm identified a list of 26 care centers where the oncology activity could grow to maximize the total accessibility in the region. These centers are either public or private hospitals, primarily located in the Avignon and Gap areas. The care centers located in high accessibility areas near Marseille and Nice

were ignored by the algorithm because improving these zones is not a priority. The care center that grew the most is Clinique Sainte Catherine, in Avignon. Interestingly, this care center was recently bought by the Unicancer group, which coordinates all the cancer centers in France. This hospital's type will change to become a new CLCC. Thus, it is expected to grow in the next years and to be equipped with more oncology services and staff.

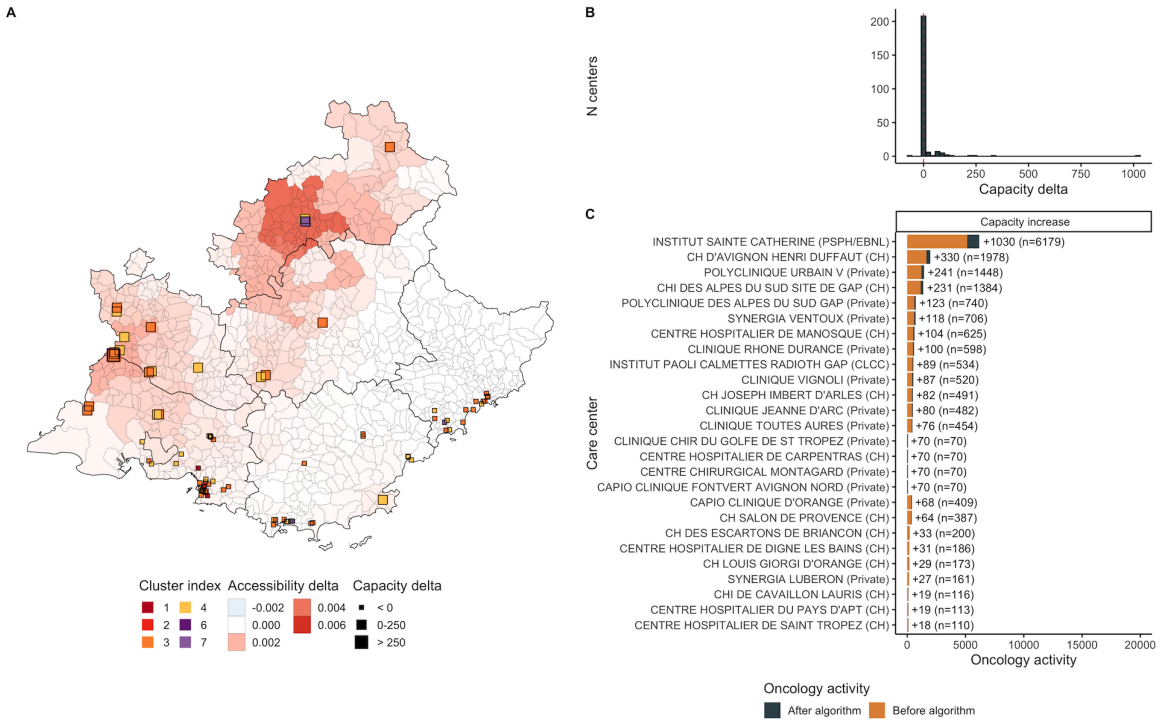


Figure 4.1: Accessibility delta in Provence-Alpes-Cote-d'Azur region after running the optimization algorithm. Map (A) displays the accessibility delta ($A_{i_{\text{after}}} - A_{i_{\text{before}}}$) by municipality. Plot (B) shows the capacity delta ($S_{u_{\text{after}}} - S_{u_{\text{before}}}$) distribution. Capacity was defined as the oncology activity: the number of patients with chemotherapy or radiotherapy and the number of medical or surgery stays related to oncology. We show the list of the care centers that grew the most (C) and by how much. For instance, the hospital Institut Sainte Catherine in Avignon, was assigned a +1,030 capacity, for a total of n=6,179. Additional activity was 3,221. 26 centers grew and 1 decreased. Median accessibility before optimization was 0.0093 and 0.0103 after, corresponding to a 11.1% increase. Accessibility increased around cities like Avignon and Gap. Care centers near Nice were left unchanged by the algorithm.

Pays de la Loire

Similarly, we describe the optimization results in the Pays-de-la-Loire region, obtained with the same parameters for the algorithm. The additional activity was 1,890. A total of 18 centers grew and 2 were decreased. The median accessibility before optimization was 0.0118 and 0.0121 after, corresponding to a 2.4% increase. Accessibility mainly grew near Le Mans, Angers and La Roche sur Yon. The three hospitals that had the highest increase in capacity were CH du Mans, a public hospital; Clinique Victor Hugo, a private hospital; and Centre Jean Bernard, a private hospital dedicated to oncology and located right next to Clinique Victor Hugo. Developing these areas will benefit patients living there and prevent them to travel to Nantes, Rennes or Angers when it is not necessary, thus avoiding long drives.

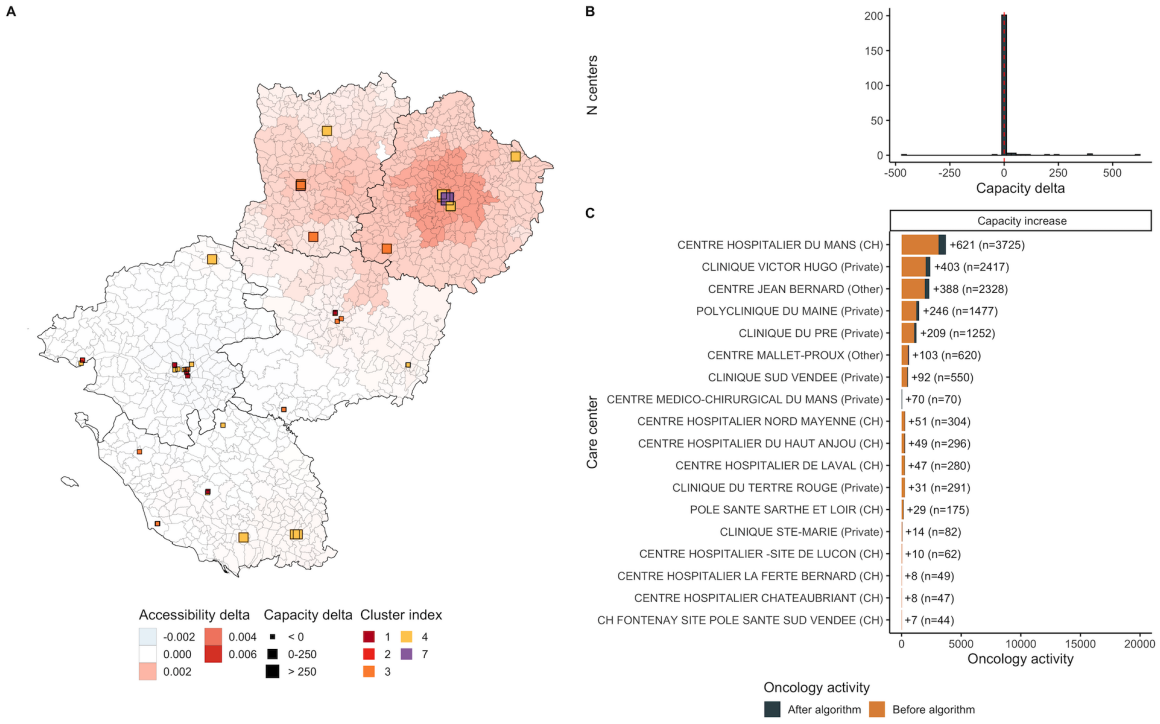


Figure 4.2: Optimization results in Pays-de-la-Loire. Additional activity was 1,890. 18 centers grew and 2 decreased. Median accessibility before optimization was 0.0118 and 0.0121 after, corresponding to a 2.4% increase. Accessibility mainly grew near Le Mans, Angers and La Roche sur Yon.

Occitanie

In Occitanie, the additional activity was 3,652. 28 centers grew and 1 decreased. Median accessibility before optimization was 0.0087 and 0.0091 after, corresponding to a 4.7% increase. Accessibility mainly grew around Perpignan, Rodez, Mende and Tarbes. The two largest cities in the region Toulouse and Montpellier were ignored by the algorithm, since this is where the largest hospitals already are. The three hospitals that were the most developed by the algorithm are Clinique Saint Pierre, a private hospital; CH Perpignan which is public and Centre Catalan d'Oncologie Saint Pierre, a private hospital dedicated to oncology located in Perpignan. The three areas picked by the algorithm are located apart from each other on the outskirts of the region.

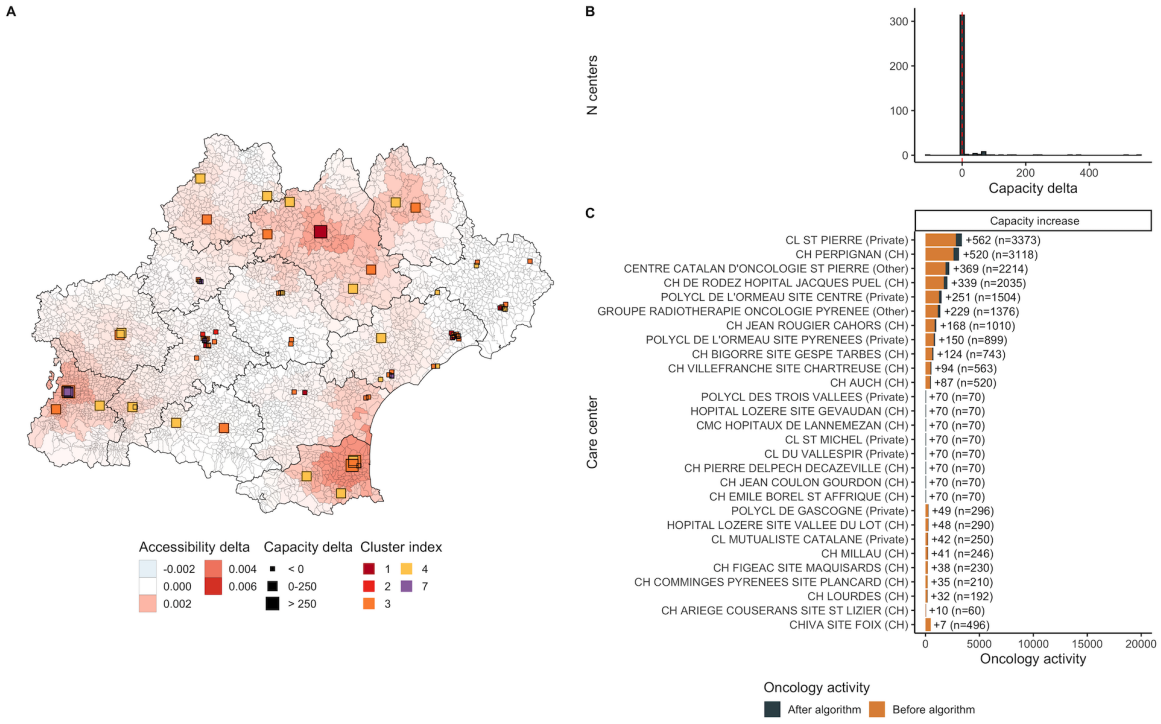


Figure 4.3: Optimization results in Occitanie. Additional activity was 3,652. 28 centers grew and 1 decreased. Median accessibility before optimization was 0.0087 and 0.0091 after, corresponding to a 4.7% increase. Accessibility grew around Perpignan, Rodez, Mende and Tarbes.

Nouvelle-Aquitaine

In Nouvelle Aquitaine, the additional activity was 3,445. A total of 25 centers grew and 1 decreased. The median accessibility before optimization was 0.0117 and 0.0119 after, corresponding to a 1.5% increase. Accessibility mainly grew around Limoges, Angoulême, and Brive-la-Gaillarde. Unlike in the Occitanie region, the algorithm picked areas that are fairly close to each others, mostly located on the northern east part of the region. The largest city Bordeaux was again left unchanged by the algorithm. The three hospitals with the largest increase in capacity are CHR/U Dupuytren in Limoges; CH Dubois in Brive-la-Gaillarde and Clinique François Chenieux in Limoges.

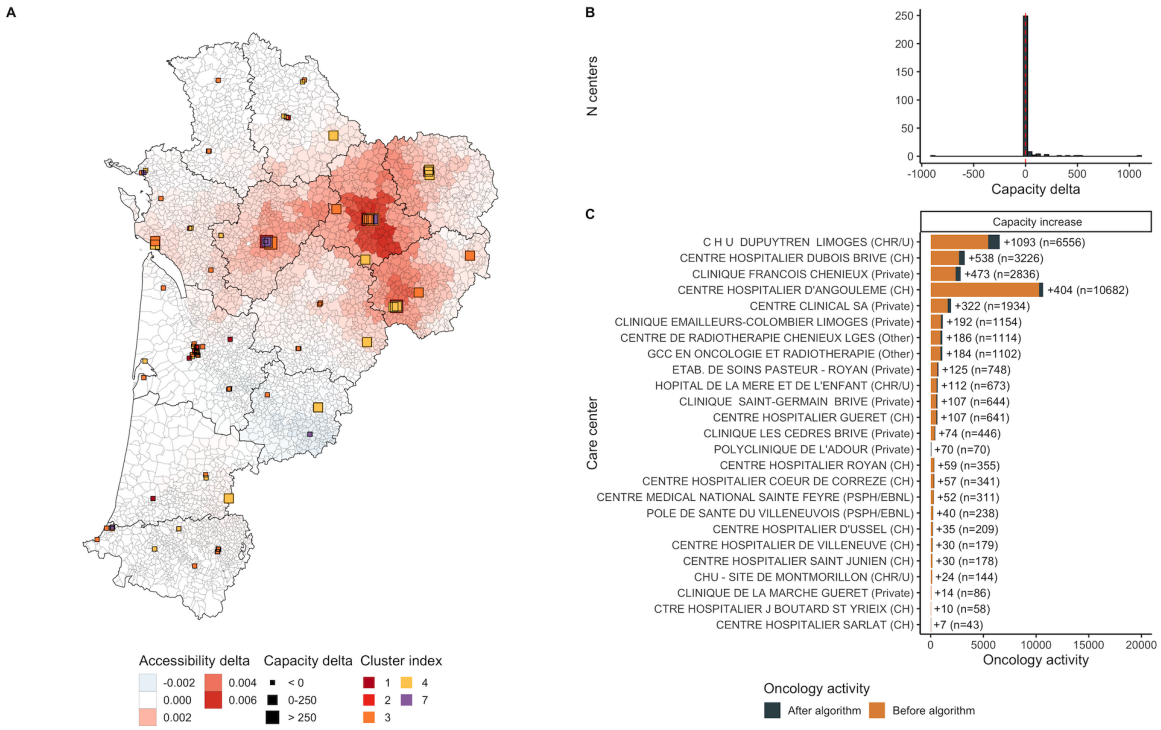


Figure 4.4: Optimization results in Nouvelle-Aquitaine. Additional activity was 3,445. 25 centers grew and 1 decreased. Median accessibility before optimization was 0.0117 and 0.0119 after, corresponding to a 1.5% increase. Accessibility grew around Limoges, Angouleme, and Brive-la-Gaillarde.

Normandie

In Normandie, the additional activity was 1,523. A total of 15 centers grew and none decreased. The median accessibility before optimization was 0.0105 and 0.0106 after, corresponding to a 1% increase. Accessibility mostly grew near Caen, Argentan, and St-Lo, on the middle-western part of the region. Hospitals near Rouen and Évreux were left unchanged by the optimization process. The algorithm mostly targeted a single hospital: the CHR/U Cote de Nacre, a public university hospital in Caen. This hospital received a +771 activity, for a total of 4628 stays. The other hospitals that were increased are smaller, and their number of stays range between 24 and 725.

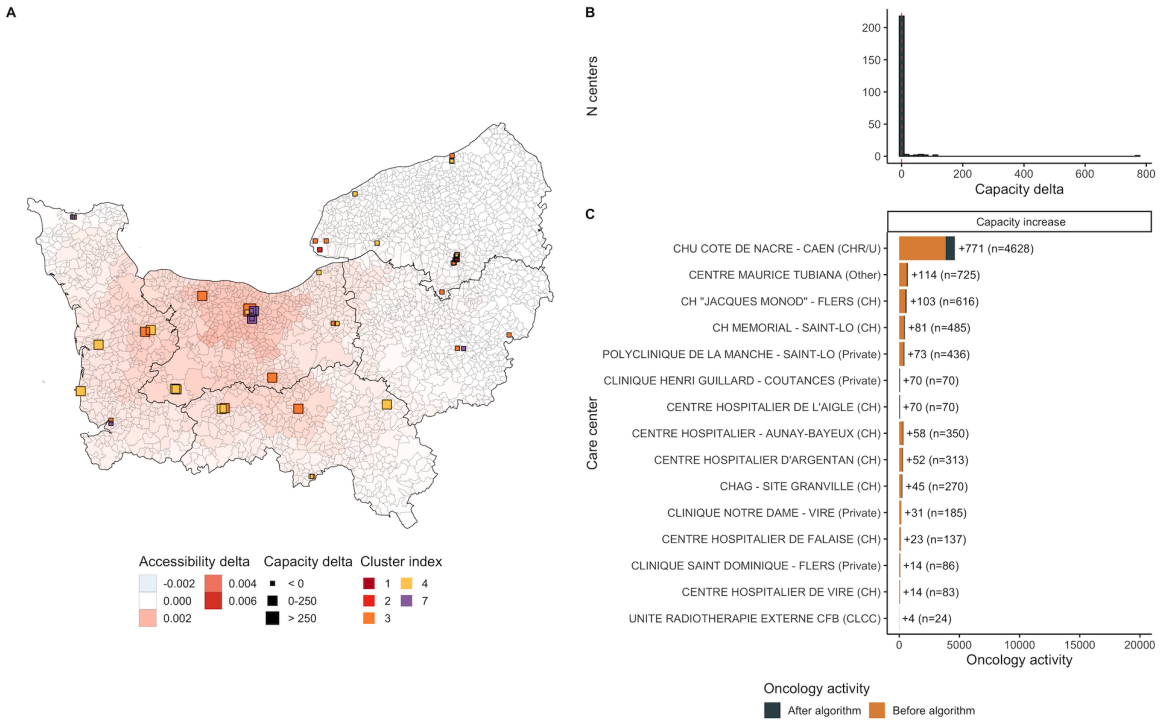


Figure 4.5: Optimization results in Normandie. Additional activity was 1,523. 15 centers grew and 0 decreased. Median accessibility before optimization was 0.0105 and 0.0106 after, corresponding to a 1% increase. Accessibility grew near Caen, Argentan, and St-Lo.

Ile-de-France

In Ile-de-France, the additional activity was 5,826. 44 centers grew and 1 decreased. The median accessibility before optimization was 0.0088 and 0.0089 after, corresponding to a 1.3% increase. Accessibility grew around Mantes-la-Jolie, Rambouillet, Melun, and Évry, on the outskirt of the region, surrounding the Paris city. Looking at the map, it is harder to distinguish specific areas that were developed, since the accessibility increase is much more spread than in the other regions. This is probably due to the relatively high population densities in the whole region. Developing the hospitals outside of the Paris city seems fair, given the tedious drive that it would take to reach the city center from the suburbs, especially due to the traffic. Moreover, the most specialized hospitals in Paris are often already saturated, from patients living in Paris or coming from other regions in the case of rare cancers.

Hauts-de-France

In Hauts-de-France, the additional activity was 2,520. A total of 29 centers grew and 1 decreased. The median accessibility before optimization was 0.01 and 0.0102 after, corresponding to a 2.1% increase. Accessibility mainly grew around St-Quentin and Valenciennes. Similarly to the results in Ile-de-France region, it is relatively hard to distinguish precise areas where the accessibility was increased, and the accessibility delta is more evenly spread around the region. However there are areas where the hospitals remained unchanged, in Lille for instance or in the southern part of the region, in the Oise department, near Beauvais, Compiègne or Senlis. The two hospitals where the capacity increase were the largest are Centre Leonard de Vinci, a private structure near Douai; CH Saint Quentin, a public hospital. Both received around +500 capacity increase, bringing them to a total of roughly 3000.

Grand Est

In Grand Est, the Additional activity was 2,663. A total of 31 centers grew and 4 decreased. The median accessibility before optimization was 0.0096 and 0.0099 after, corresponding to a 3% increase. Accessibility grew around Troyes and Épinal. In this region, the municipali-

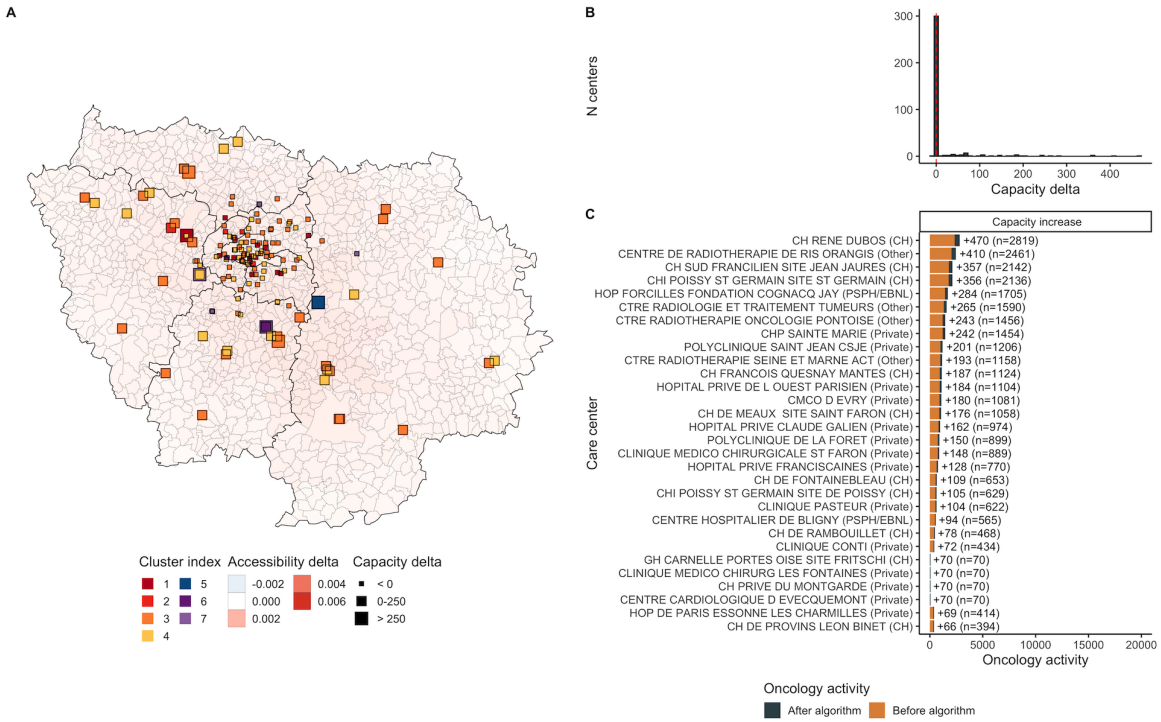


Figure 4.6: Optimization results in Ile-de-France. Additional activity was 5,826. 44 centers grew and 1 decreased. Median accessibility before optimization was 0.0088 and 0.0089 after, corresponding to a 1.3% increase. Accessibility grew around Mantes-la-Jolie, Rambouillet, Melun, and Évry.

ties with the largest accessibility scores are located on the eastern end, along the Germany frontier. The hospitals in these areas were not increased, and the algorithm focused on sub-urban cities like Troyes, where patients are sometimes traveling all the way to Paris for certain types of cancer. Among the grown hospitals, the first two ones are in Troyes. Clinique de Champagne, a private structure, received an additional activity of 543, totalling 3260 stays. Next, the CH Troyes, a public hospital received 375 additional stays, bringing the capacity to 2248.

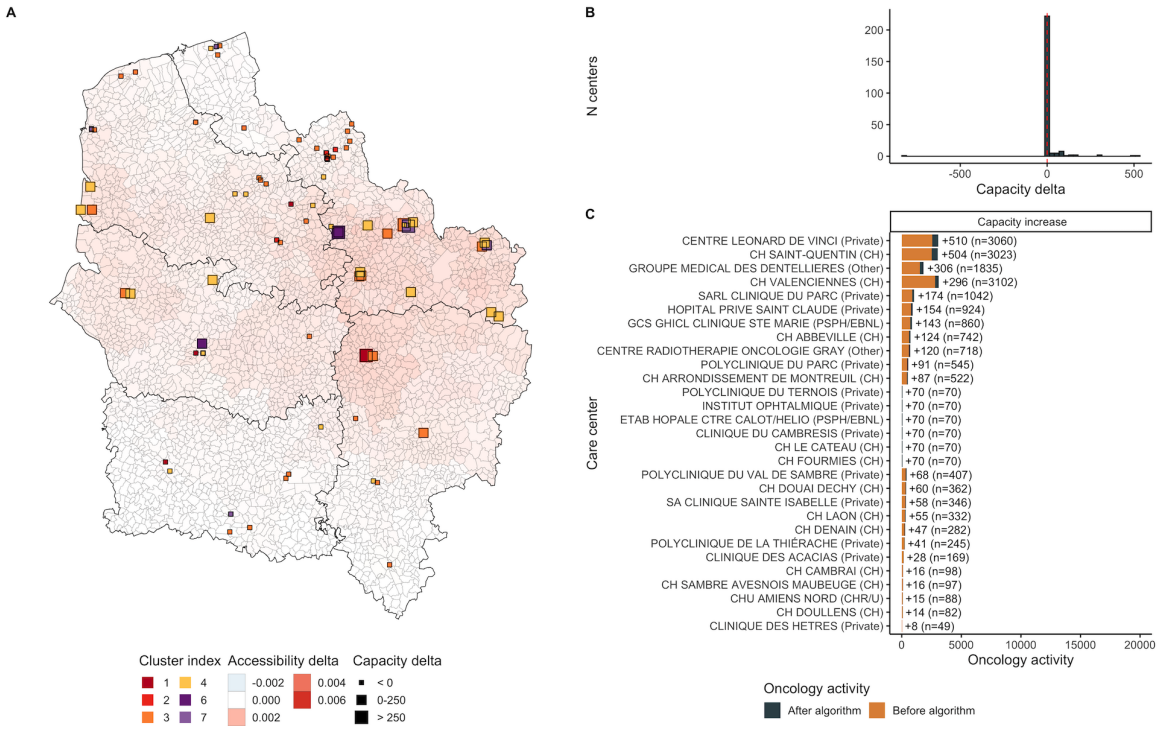


Figure 4.7: Optimization results in Hauts-de-France. Additional activity was 2,520. 29 centers grew and 1 decreased. Median accessibility before optimization was 0.01 and 0.0102 after, corresponding to a 2.1% increase. Accessibility grew around St-Quentin and Valenciennes.

Centre-Val de Loire

In Centre-Val de Loire, the additional activity was 1,072. A total of 10 centers grew and 1 decreased. The median accessibility before optimization was 0.0099 and 0.0102 after, corresponding to a 2.9% increase. Accessibility grew around Bourges and Châteauroux areas. Compared to the other regions, very few hospitals were increased by the algorithm. Two departments were affected by the modifications: Indre and Cher. Cities like Tours, Chartres and Orleans were left as such by the optimization process. The two hospitals that were the most affected by the algorithm are private structures, that each received around 200 capacity increase.

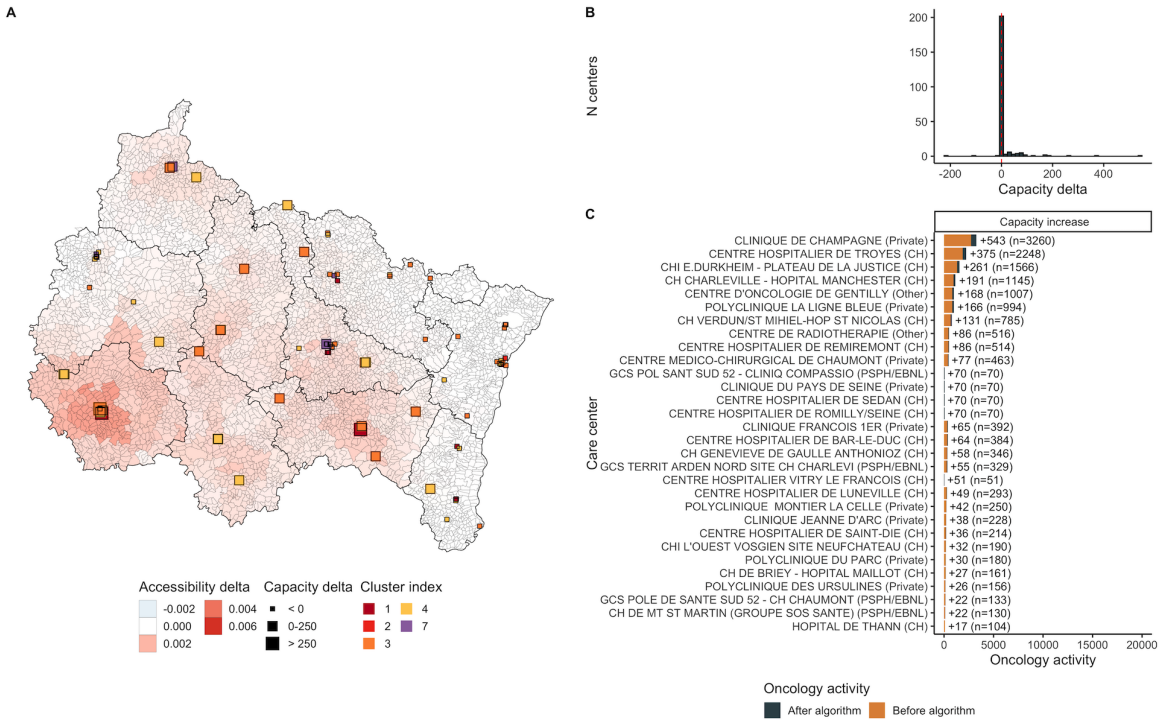


Figure 4.8: Optimization results in Grand-Est. Additional activity was 2,663. 31 centers grew and 4 decreased. Median accessibility before optimization was 0.0096 and 0.0099 after, corresponding to a 3% increase. Accessibility grew around Troyes and Épinal.

Bretagne

In Bretagne, the additional activity was 1,773. A total of 10 centers grew and 2 decreased. The median accessibility before optimization was 0.0131 and 0.0134 after, corresponding to a 2.4% increase. The accessibility grew mostly around St-Brieuc area, in the northern part of the region. The accessibility distribution in Bretagne was among the highest and most homogeneous compared to the other regions. Accessibility was lower in the inland areas between Cotes d'Armor and Morbihan departments. This optimization affected mostly the Cotes d'Armor department, but spread near the frontier of Morbihan. Since the population density is lower in the center part of the region, the algorithm focused on the hospitals located in the sub-urban areas near Saint Brieuc. We pointed earlier that even though the accessibility in the inland part of the region was not as high as on the outskirts, the patients

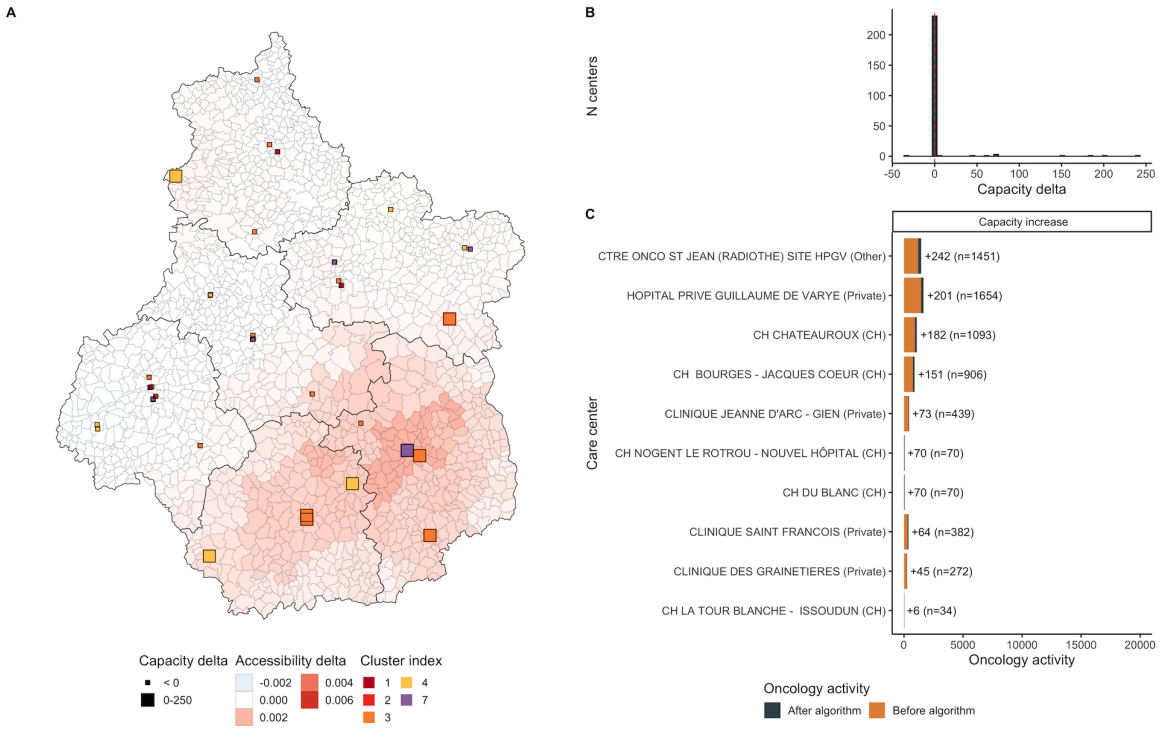


Figure 4.9: Optimization results in Centre-Val-de-Loire. Additional activity was 1,072. 10 centers grew and 1 decreased. Median accessibility before optimization was 0.0099 and 0.0102 after, corresponding to a 2.9% increase. Accessibility grew around Bourges and Châteauroux.

average travel duration remained relatively low, limiting the travel burden.

Bourgogne-Franche-Comte

In Bourgogne-Franche-Comte, the Additional activity was 1,330. A total of 13 centers grew while none decreased. The Median accessibility before optimization was 0.0096 and 0.0098 after, corresponding to a 1.9% increase. Accessibility grew around Nevers and Auxerre, on the western side of the region. The largest hospitals in this region are located near Dijon, the largest city. The hospitals in this area were left unchanged. The optimization had the largest effect on departments like Nièvre and Yonne. The number of modified hospitals is relatively low compared to the other regions, and the largest additional capacities are in the

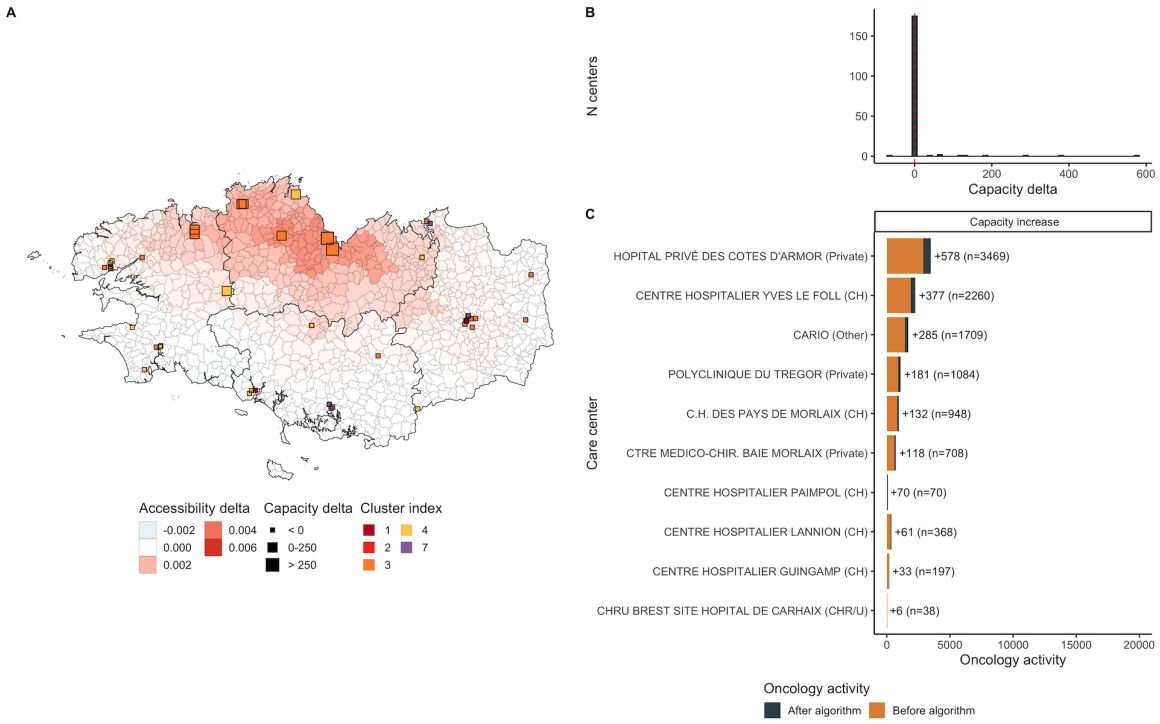


Figure 4.10: Optimization results in Bretagne. Additional activity was 1,773. 10 centers grew and 2 decreased. Median accessibility before optimization was 0.0131 and 0.0134 after, corresponding to a 2.4% increase. Accessibility grew around St-Brieuc.

hundreds. Among the most affected hospitals, two of them are located in Auxerre: the CH Auxerre which is a public structure; and the Centre de Radiothérapie du Parc, dedicated to radiotherapy.

Auvergne-Rhone-Alpes

In Auvergne-Rhone-Alpes, the additional activity was 3,883. A total of 23 centers grew and 2 decreased. The median accessibility before optimization was 0.0092 and 0.0095 after, corresponding to a 3.2% increase. Accessibility grew around Moulins, Montluçon, Le Puy en Velay, Clermont-Ferrand and Aurillac. The departments that were the most affected by the capacity increase were Allier, Puy de Dome, Cantal and Haute Loire, mainly on the eastern part of the region. The hospitals near Lyon in Rhone department were left unchanged. De-

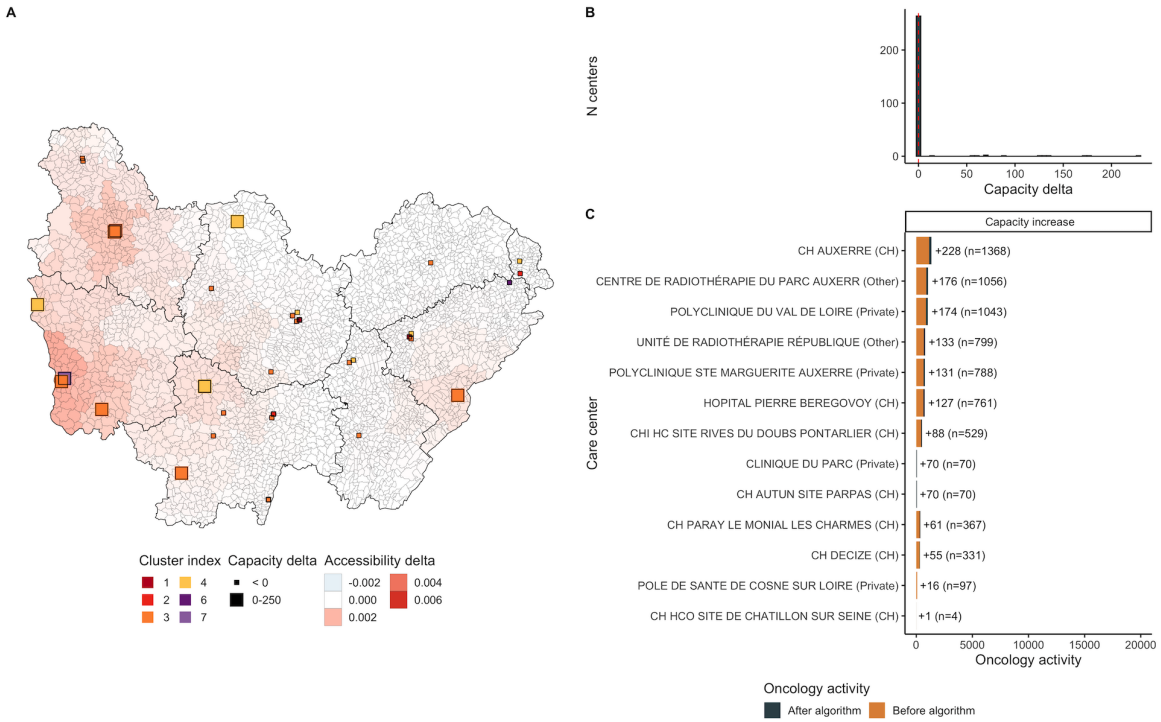


Figure 4.11: Optimization results in Bourgogne-Franche-Comté. Additional activity was 1,330. 13 centers grew and 0 decreased. Median accessibility before optimization was 0.0096 and 0.0098 after, corresponding to a 1.9% increase. Accessibility grew around Nevers, Belfort, Vesoul and Auxerre.

veloping the hospitals in the eastern part makes sense to avoid patients traveling from there to Lyon. Indeed, the routes can be tedious, and the drives relatively long. The hospital that was developed the most is CLCC Jean Perrin in Clermont-Ferrand, a large oncology dedicated hospital. The algorithm allocated an additional capacity of 1,171, bringing the hospital to a 7,024 overall activity. In this region, we notice that multiple hospitals from the cluster 1 were developed. In other regions, such hospitals are most of the time ignored by the algorithm, and structures from cluster 3 or 4 are often developed instead. Developing hospitals from cluster 1 might be easier than developing hospitals from least specialized clusters because they have more oncology services that are already established. However, we must make sure they are not already saturated, due to the higher oncology volumes they usually operate at.

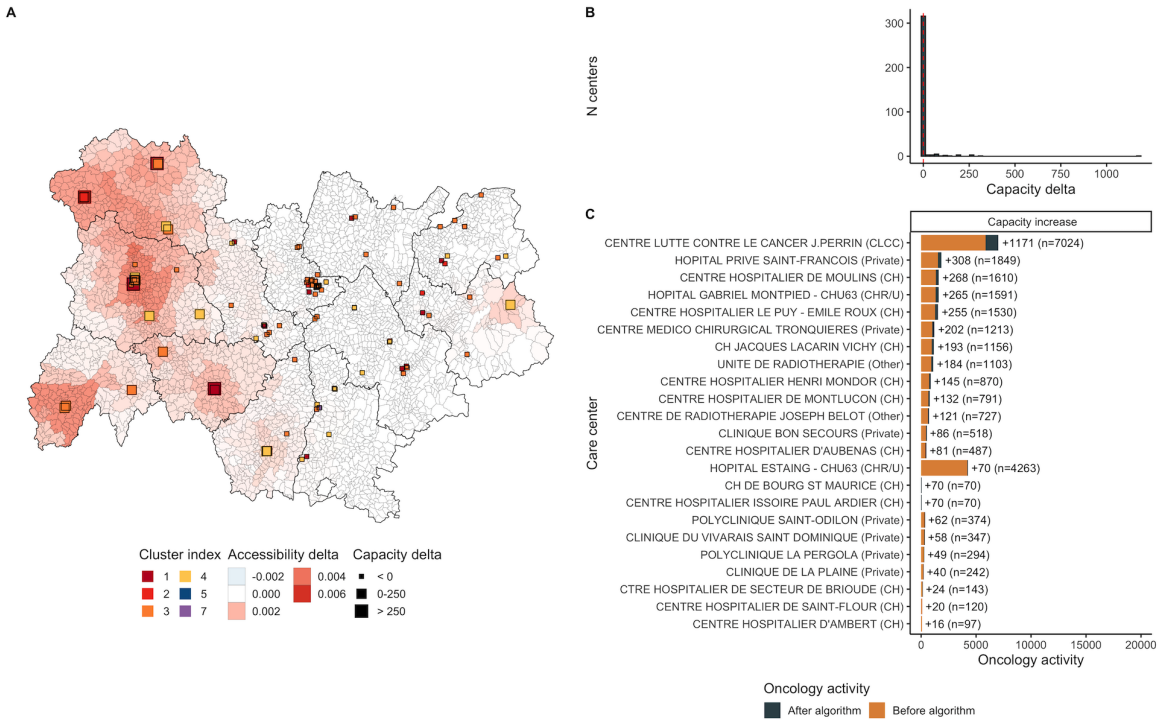


Figure 4.12: Optimization results in Auvergne-Rhone-Alpes. Additional activity was 3,883. 23 centers grew and 2 decreased. Median accessibility before optimization was 0.0092 and 0.0095 after, corresponding to a 3.2% increase. Accessibility grew around Moulins, Montluçon, Le Puy en Velay, Clermont-Ferrand and Aurillac.

4.2.2 Oncology Accessibility web application

We created a web application to run the optimization algorithm with the desired hyper parameters, for any region in metropolitan France. It is a two-screen web application. The first screen is the home page that shows a summary of the project, some descriptive statistics about the regions and care centers in metropolitan France and a form that lets the user pick the optimization hyper parameters. The next screen displays the optimization results, with an interactive map, distribution plots and a table with the list of affected care centers in the region. We now detail the list of the optimization hyper parameters available in the form. First, the user can pick which region to run the algorithm on. The supply variable can also be chosen. It defaults to oncology activity, but the user can also choose more specific variables

like medical and surgery oncology, radiotherapy, or chemotherapy activity. Then, the additional capacity, maximum growth and decrease percentage are also editable. Finally, fine tuning based on the clusters is possible. We can set the maximum capacity of the least specialized cluster, the maximum decrease of the highly specialized clusters and the maximum capacity of care centers in intermediate clusters and without initial activity. We developed the application using python programming language and Flask micro-framework. We used the plotly and folium librairies for drawing the plots and maps. All these technologies are free and open source.

A form, displayed on Figure 4.13, allows the users to choose the parameters for the optimization algorithm. The form fields are:

- Region: The region where the optimization will be run on. The optimization is ran on the whole metropolitan France to avoid border effect. However, care centers that are not from the given region are not allowed to grow/decrease. Only the care centers and municipalities from the given region and the surrounding departments are displayed.
- Supply variable: The variable to use as capacity for the accessibility score. This is the value that will encode supply, balanced the population demand. We let the user chose from multiple variables, to make sure different needs could be covered. The variable choices so far are:
  - Oncology activity: The supply variable equals the number of medical or surgery stays related to cancer + the number of chemotherapy and radiotherapy patients. This is the default variable, which was used in the previous methods and results.
  - MCO activity: The supply variable is the number of medicine, surgery and obstetric stays. With this supply variable, we no longer focus on oncology accessibility only. The accessibility score is more global and should be interpreted more carefully.
  - Chemotherapy activity: The supply variable in this case is the number of chemotherapy patients per facility.

  - Radiotherapy activity: The supply variable is now equal to the number of radiotherapy patients in the hospital.
  - Oncology medical and surgery activity: This indicator is the number of medical or surgery stays related to cancer. It is equal to the oncology activity without chemotherapy and radiotherapy patients.

- Additional supply: The activity to be added to the current overall activity. Setting this parameter to 0 will lead to an optimization constraint with "constant" activity, meaning that a care center will have to decrease to let another one grow. If this number is set between 0 and 1, the corresponding percentage of the current activity is added. e.g: 0.03 will add 3% of the current activity.
- Max growth percentage: The maximum growth percentage of a care center. If set to 20%, the care center will not be allowed to grow by more of 20% of its current activity.
- Max decrease percentage: The maximum decrease percentage of a care center. If set to 20%, the care center will not be allowed to decrease by more of 20% of its current activity. If set to 0, the care centers activity can't decrease.
- Low cluster max capacity: The maximum capacity that the care centers from the least specialized cluster can reach. If set to 0, these care centers can't receive any activity and will be emptied if they originally had some.
- High cluster max decrease: This is similar to the "max decrease percentage" parameter, but only applied to the care centers from the most specialized cluster. If set to 0, these care centers won't be allowed to decrease.
- Maximum new capacity: The maximum capacity that the care centers with 0 activity can receive, unless they are within the least specialized cluster. In this case, this parameter will be ignored and "low cluster max capacity" will be used.

Once the parameters are set, the optimization algorithm runs and displays the results on an interactive map, as shown on Figure 4.14. The accessibility delta is displayed by default

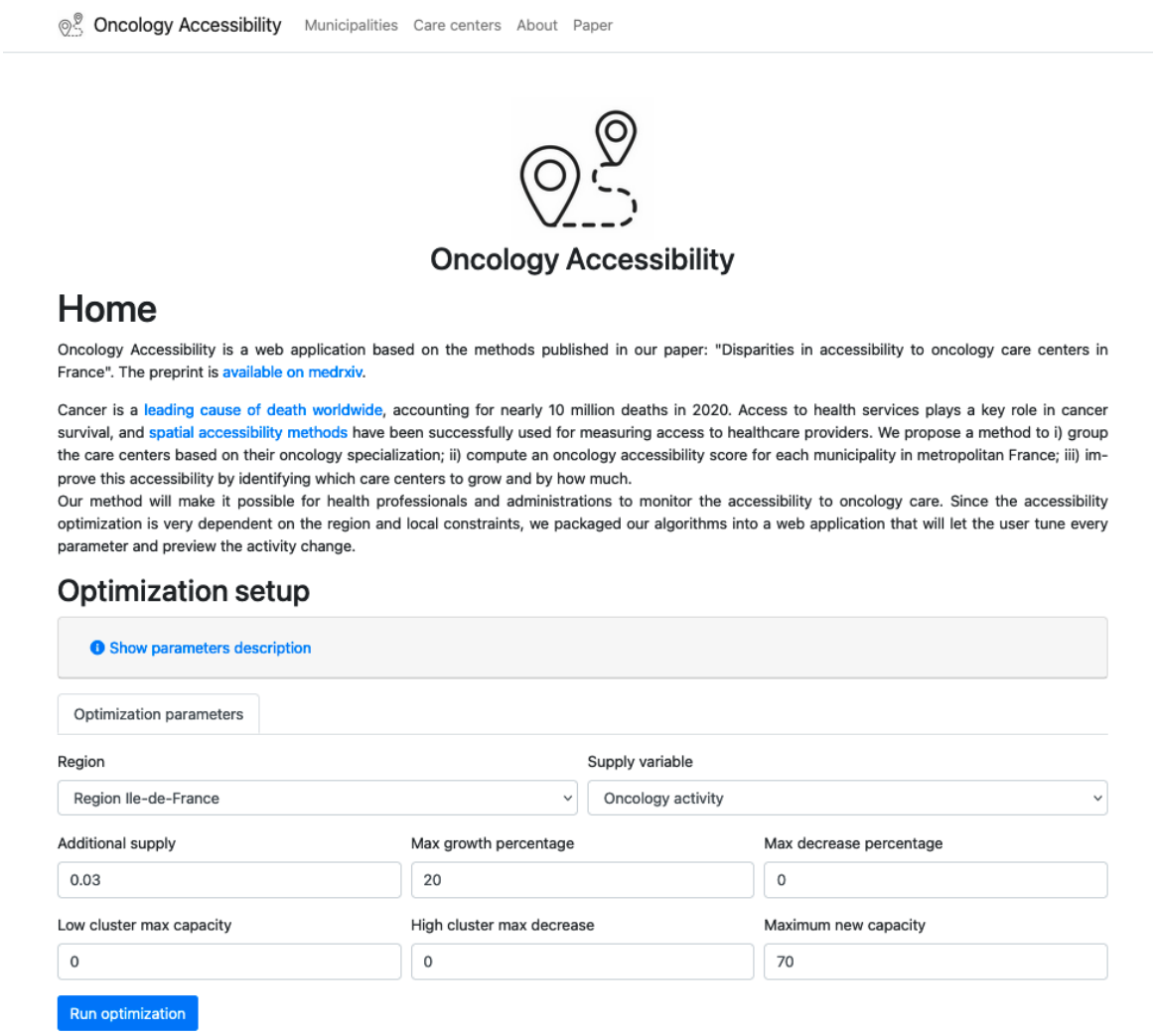


Figure 4.13: Oncology Accessibility: Homepage.

on the map, as well as the hospitals colored in green, red or green whether the hospital was grown, decreased or remained as is. The accessibility delta is defined as the difference between the municipality accessibility after optimization and before optimization. A positive delta means that the municipality accessibility increased due to grown hospitals nearby. A high delta is displayed in red on the map whereas a municipality with a null delta is colored in pale yellow. On this interactive map, we can also display municipalities population densities, as well as the accessibility before and after optimization. Three histograms are placed below the map, to compare the municipalities accessibility distribution before and

after optimization, as well as the accessibility delta distribution.

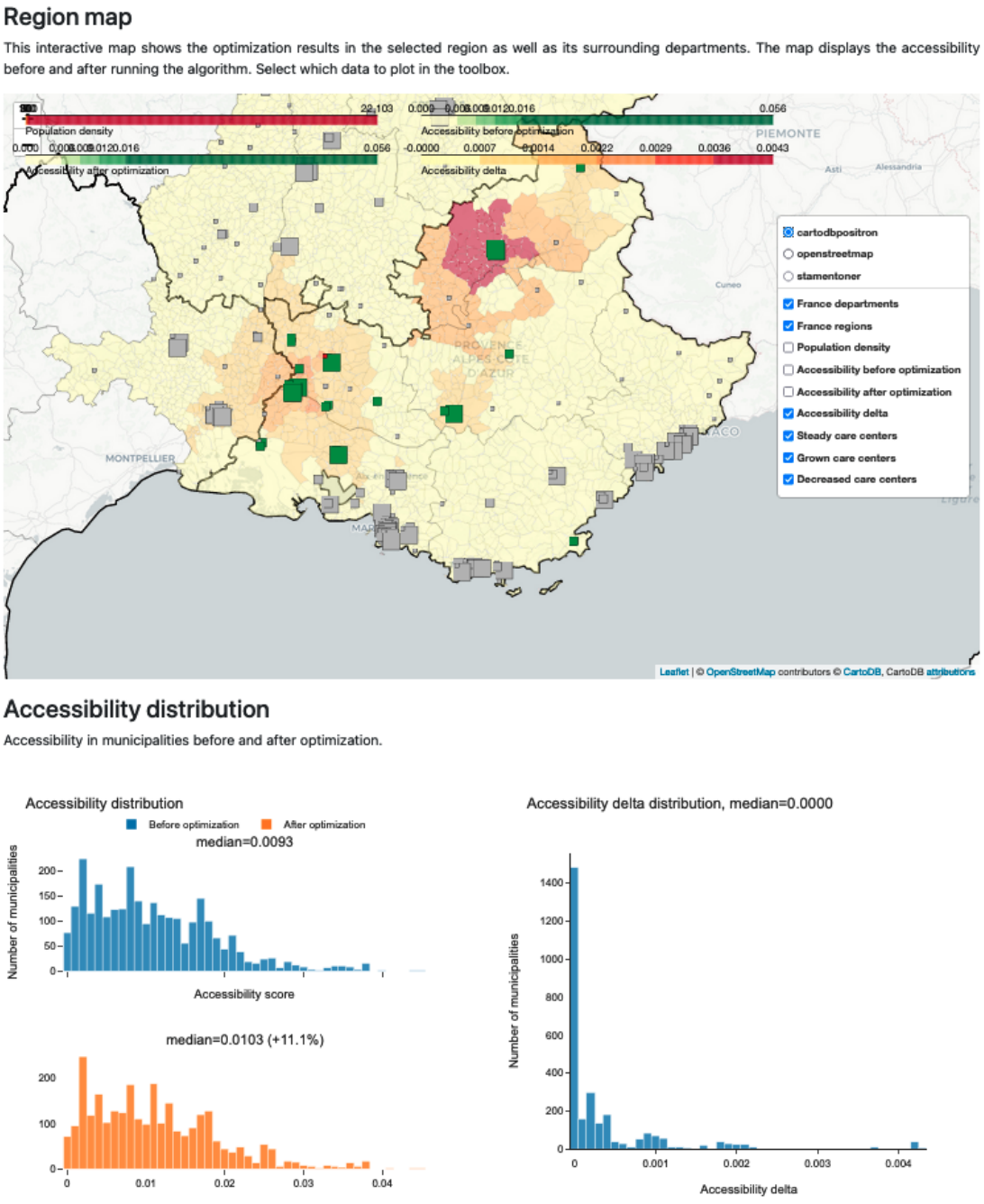


Figure 4.14: Oncology Accessibility: Optimization results. The accessibility delta is displayed by default on the map, as well as the hospitals colored in green, red or green whether the hospital was grown, decreased or remained as is.

Finally, the list of hospitals in the region is displayed below the histograms. The orignal capacity and modified capacity are shown, as well as the percentage of increase. The hospital category and cluster are displayed, for better interpretation of the modifications. On a separate web page, it is possible to get the list of municipalities and their accessibility scores, for every region. An interactive map is also displayed below the table.

4.2.3 Open source code: application on the New York City hospitals

We open sourced the code for accessibility computation and CAMION algorithm. The code is available in the following Github repository: https://github.com/ericdaat/CAMION. As an example to showcase our package, we applied our method to Health Facilities in New York City. We used datasets downloaded from NYC Open Data website, which lists free public data from New York City agencies and other partners. We downloaded the Zip Codes boundaries and census statistics in New York City, provided by the Department of Information Technology and Telecommunications. We retrieved the list of health facilities in the New York State, as well as their certifications for services and beds. Both datasets were provided by the New York State Department of Health. We only kept the health facilities located in New York City, with Medical / Surgical beds. Every hospital has Latitude / Longitude coordinates. We used Zip Codes polygons centroids as reference point to compute the travel between Zip Codes and hospitals. We used the Zip code population as P_i, to encode the demand variable. The supply variable S_u was the number of Medical / Surgery bed for each Health Facility u. We used the geodesic (straight) distance between health facilities coordinates, and Zip Code centroid coordinates as distance matrix. The previously cited datasets can be downloaded on the following links: Zip Codes boundaries and census statistics in New York City; List of health facilities in NY State; Health facilities certifications for services and beds.

In the following paragraphs, we describe the hospitals in New York City and provide code snippets to run our method. We then display the accessibility scores and optimization results, similarly to what we did earlier for the different regions in metropolitan France. We stress that the healthcare management is very different in the US and in France, so we do

not claim that our results should be applied as such in New York City. We only used this dataset because it was public, and we could easily replicate our method to show a working example to interested users.

We first showed the spatial distribution of the included health facilities in New York City, as seen on Figure 4.15. Map (A) illustrates the New York City zip codes, colored by population. The facilities are drawn as pictograms, sized by the number of beds and colored by the county to which they belong. The number of beds distribution is shown on histogram (B), and the facilities with the most beds are listed on barplot (C). The largest hospitals are mostly located in New York county.

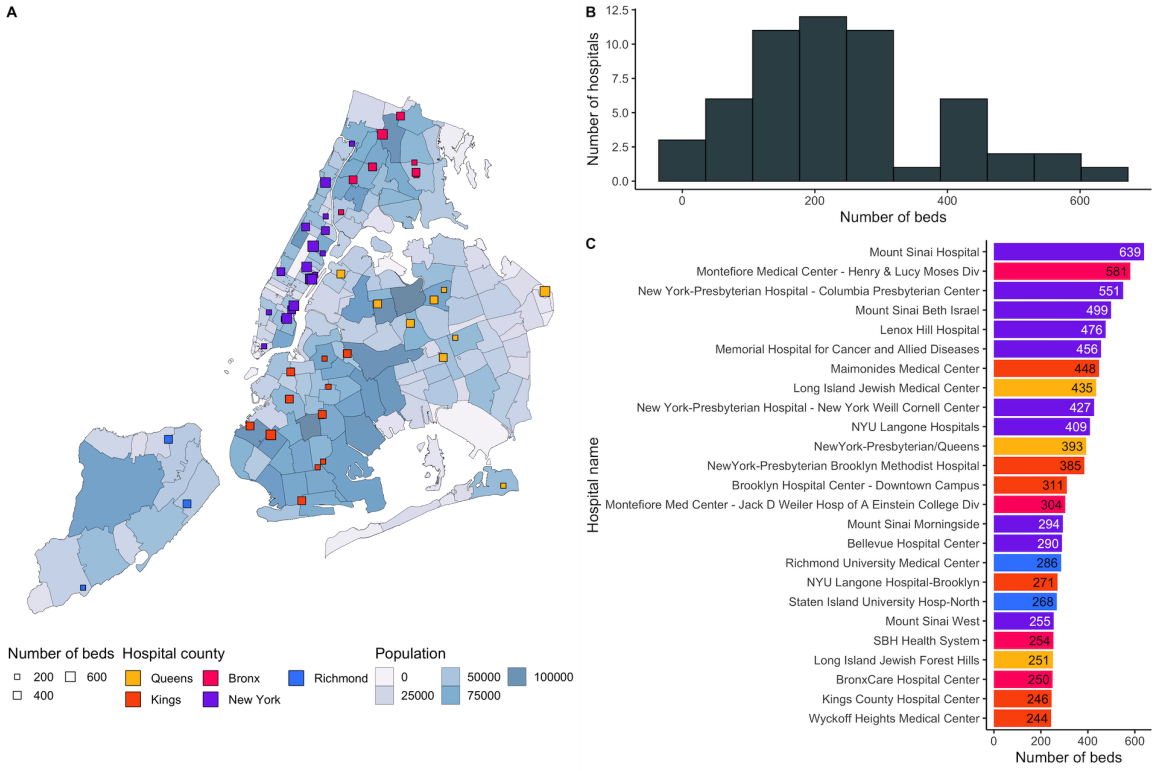


Figure 4.15: Health facilities with Medical / Surgery beds in New York City. We included 55 facilities with a total of 13,443 beds. Map (A) shows the geographical location of the facilities, colored by county, and sized by number of beds. The distribution of the number of beds is shown on (B). The top 30 facilities with the highest number of beds are listed on (C) and colored by county. The largest facilities are in New-York County.

We now show how to use our package to compute the accessibility scores with the E2SFCA algorithm. For clarity, we used randomly generated data, but a working example on the New York hospitals is avaiable on the Github repository: https://github.com/ericdaat/CAMION/blob/main/paper/methods.ipynb. For this example we sampled 100 facilities and 10 population locations. The travel impedance were also sampled, with values between 1 and 100. The impedance weights have been set similarly to our paper, with distance bins of 30, 60 and 90. Hence the maximum catchment area was set to 90. The following code snippet illustrates how to initialize the data and run the algorithm.

Listing 4.1: Compute accessibility score with E2SFCA

```
from camion.fca import E2SFCA

# Declare variables
P_i = np.random.rand(100) # Facilities
S_j = np.random.rand(10) # Population locations
D_ij = np.random.randint(low=1, high=100, size=(100, 10)) # Travel impedance

# Init E2SFCA algorithm
e2sfca = E2SFCA(
    S_j=S_j,
    P_i=P_i,
    D_ij=D_ij
)

# Choose weights for travel impedance
weights = [(30, 1), (60, 0.42), (90, 0.09)]

# Compute accessibility scores
A_i = e2sfca.compute_accessibility_score(weights)
```

In this paragraph, we describe the accessibility results that we obtained on the New York

City dataset. The results are illustrated on Figure 4.16. The accessibility scores are displayed on map (A) for every zip code in the city. Since the largest hospitals were located in the New York county, it is no surprise that the highest accessibility scores are located in that area. The Richmond county seems to have the lower accessibility values, as shown on boxplot (C). The histogram (B) shows the accessibility distribution. We see that the majority of the zip codes in New York City have a high accessibility score. Finally, scatter plot (D) compares the accessibility scores with the population for every zip code. There does not seem to be a correlation between both series, as even zip codes with lower population can have a good accessibility, especially in the New York county.

After computing the accessibility scores, we are now interested in running the optimization algorithm. As we did previously, we first show a code snippet on the previously randomly generated data, and then we display our results obtained on the New York City dataset. In the following code snippet, we first define the optimization parameters, namely the budget and the maximum growth percentage for every facility. In this case, we picked a budget of 1,000 of beds, that will be spread between the 10 facilities. The growth percentage is set to 30%, meaning that no facility can grow more than 30% of its current capacity. We then initialize the optimization algorithm, which could either be overall optimization or maxi-min. Both methods have similar code expressions.

Listing 4.2: Optimize accessibility with CAMION

```
from camion.optimization import RegularOptimizer, MaxiMinOptimizer

# Define optimization parameters
budget = 1000
growth_percentage = 0.3

# Init regular optimizer
regular_optimizer = RegularOptimizer()

# Init maximin optimizer
maximin_optimizer = MaxiMinOptimizer()

# Run optimization with the optimization parameters
S_j_new_regular = regular_optimizer.run_optimization(
    S_j, P_i, W_ij, budget, growth_percentage
)
S_j_new_maximin = maximin_optimizer.run_optimization(
    S_j, P_i, W_ij, budget, growth_percentage
)
```

The Figure 4.17 displays the optimization results on the New York City dataset, using both methods, namely overall optimization (A) and maxi-min (B). The differences between the two optimization strategies are clearly visible. The overall optimization maximizes efficiency, thus the algorithm focuses on areas where the population is higher, like New York or Kings counties. On the contrary, the maxi-min approach focuses on equity, and will address the areas with low accessibility scores first, like Richmond county for instance.

4.3 Conclusion

In this chapter, we introduced CAMION, an optimization algorithm based on Linear Programming (LP) to optimize the accessibility distribution. The accessibility was computed with the Enhanced Two Step Floating Catchment Area (E2SFCA) algorithm as seen previously, but our method can generate to more Floating Catchment Area derivatives. We introduced two optimization strategies, that either maximizes efficiency or equity in the accessibility distribution. When we applied our method in metropolitan France, we chose to optimize for efficiency and the optimization task was to maximize the total accessibility instead of the minimum value. We ran the algorithm on every region in metropolitan France and displayed the results on static maps. However, we believe that our method could have larger benefits if the users could run the algorithm themselves with the parameters they judge best. For this reason, we developed "oncology-accessibility", a web application that embeds our results and methods to let the users interact with our optimization algorithm and visualize the results on interactive maps and figures. This way, several optimization strategies could be tested to find the best approach to reduce disparities in accessibility to oncology care in the country. Looking at the optimization results for every region, we observed two types of optimization outcome. For most regions, the algorithm manages to find a couple of areas where the accessibility can be locally improved, like it did in Provence-Alpes-Cote-d'Azur near Gap and Avignon. However, for regions like Ile-de-France and Haut-de-France, the hospital capacity increase is more uniformly distributed across the region. Most of the time, the algorithm left untouched the large care centers located in dense cities with good accessibility. This can be explained by the relatively low value of the additional activity parameter: with a very large value of additional activity, every care center will grow. If we keep it low, the algorithm identifies in which areas hospital capacity should be increased in priority. The quality of oncology care is linked with the care centers' volume. A care center with a very low activity is less likely to provide decent care. As a result, INCA defined several thresholds that forbid care centers with very low activity to keep operating. Similarly, the care quality in a saturated

care center won't be good either, since patients are more likely to wait longer before diagnosis or between interventions. While it is easy to spot care centers with low activity, it is harder to judge if a care center is over-crowded, and we should be careful when attributing new activity to the hospitals. We based the 20% max growth out of the previous centers' activity increase. This percentage could be tailored to the center cluster or current activity. Volume is not the only factor determining care quality. More sophisticated indicators like average delay between diagnosis and first treatment can tell whether a care center is in line with the care pathways recommendations. Care centers with activities lower than the thresholds, or with a large proportion of degraded pathways should be handled with care by our algorithm. Accessibility optimization depends on many factors and healthcare professionals will not have the same uses for our algorithm. Some may consider that for a care center to grow another should decline, where others would rather not decrease any centers' activities. Moreover, the healthcare planning is very different from a region to another, and even within the regions departments are showing disparities. Hence, we cannot expect the algorithm to be used with the same parameters on every region. For all these reasons, we believe that providing a web application to run the algorithm and choose the parameters is the most useful way to the help healthcare professionals improve the current situation.

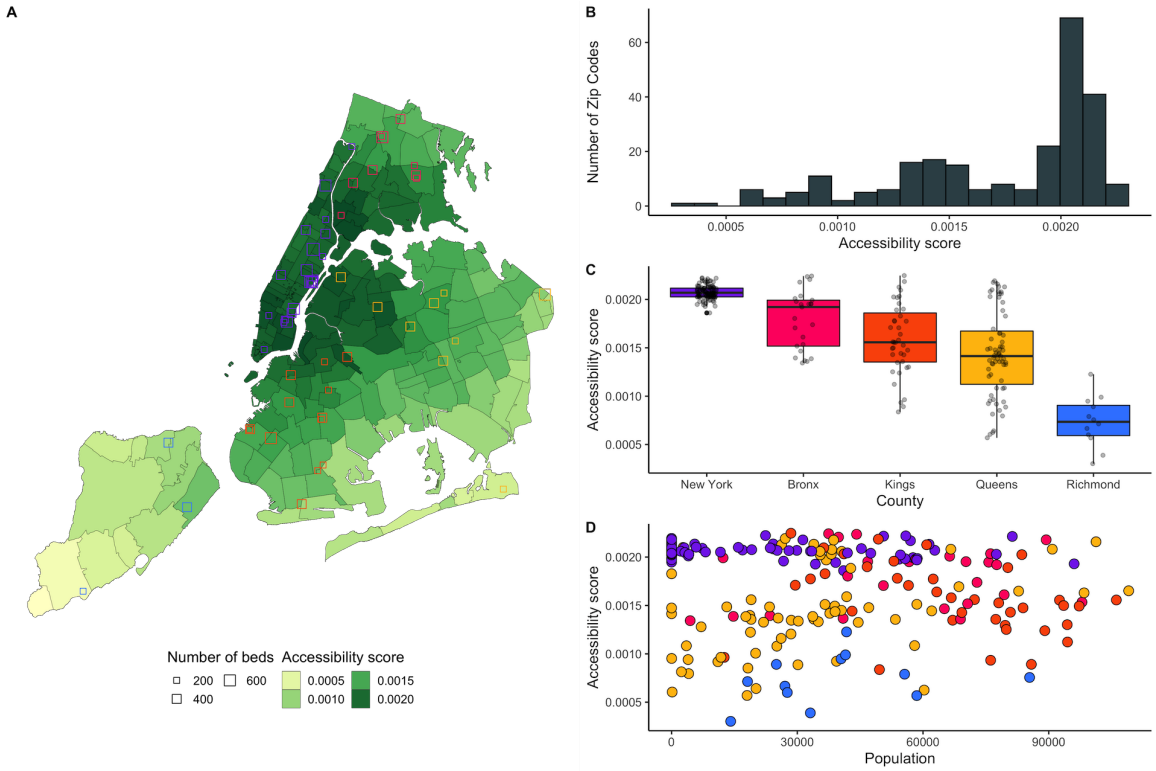


Figure 4.16: Accessibility to Medical / Surgery beds in New York City. Accessibility score was computed with the Enhanced Two Step Floating Catchment Area method, with a 45 km maximum catchment area. The geographical distribution of the accessibility score is shown on map (A). Zip codes are colored by accessibility score. Facilities are sized by number of beds and colored by county. The overall accessibility distribution is shown on (B). New-York County has the highest accessibility distribution where Richmond has the lowest (C). Accessibility seems to be higher in dense areas but there is no significant correlation between accessibility and population (D).

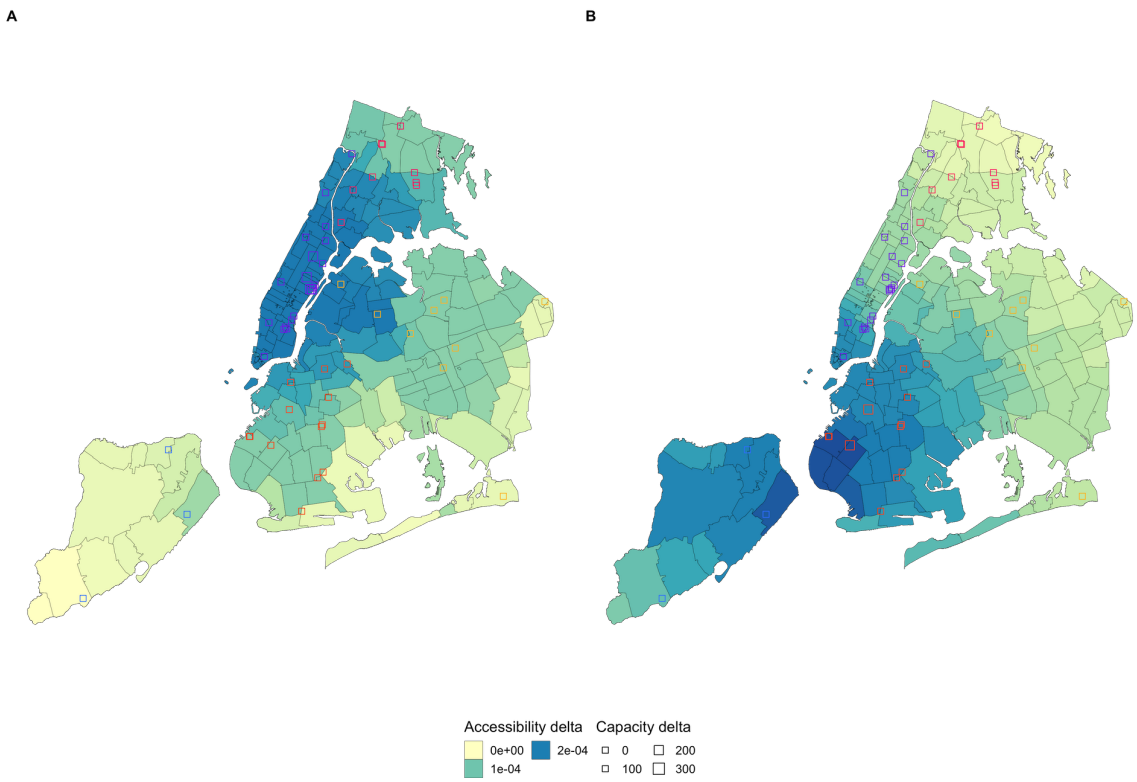


Figure 4.17: Accessibility delta after running the optimization algorithm. Both overall and maxi-min optimization algorithms are run. The optimization results are illustrated on maps (A) and (B) respectively. We displayed the accessibility delta as the difference of accessibility after and before the optimization. Every zip code is colored by accessibility delta. The health facilities are displayed as squares, sized accordingly to the capacity increase. The overall optimization increased facilities around New-York and Queens Counties (A). The maxi-min algorithm targeted Richmond facilities in priority (B).

Chapter 5

Optimizing patients travel

This chapter will be part of a research article, currently being written.

5.1 Methods

5.1.1 Travel burden index

In this section, we detail our method for computing the travel burden score. We used the PMSI database to identify which hospitals were the patients visiting from their population locations. We kept population locations and hospitals located in metropolitan France only. From these pairs, we retrieved routes from the Mapbox Directions API, with population locations as starting point and hospitals as destinations. We used driving car as the default mean of transportation since most patients travel with personal car or taxi to the hospital. The Mapbox API returns an array of routes ordered by descending recommendation rank. We kept the first route for our analysis. From this route, the overall duration and distance were returned directly by the API. Addition-ally, we extracted more variables: the number of roundabouts and the road sinuosity. The road sinuosity was computed as the ratio between the GPS distance and straight distance. The sinuosity is 1 for perfectly straight roads and increases with the number of turns. We computed this ratio for every road leg and summed them up to obtain the overall road sinuosity. We apply standard scaling (0 mean, unit vari-

ance) on these 4 variables, and we ran a PCA on top of the scaled data. We used the first PCA component as our score.

5.1.2 Carbon footprint estimation

We now explain how we estimated the CO_2 emissions from a driving route. We only consider the direct emissions, proportional to the traveled distance and car fuel consumption. As mentioned earlier, we extracted the GPS routes between population locations and hospitals. For each pair of locations, we have the number of patients and number of individual stays. We use the number of stays as number of travels between population locations and hospitals. We stored the overall distance extracted from the Mapbox API for each route. However, we do not know which car was used by patients during their visit to the hospital. Instead, the average CO_2 emission rate obtained from the French Agency for the Environment and Energy Management (ADEME) to estimate the emissions. Emissions were computed for every pair of population locations and hospitals, as the product between the number of patients stays, the GPS distance and the average CO_2 emission rate. In 2018, the average emission rate was 112 grams of CO_2 per kilometer. We should mention that the 2018 average emission rate is calculated from the new cars sold that year. The average emission rates for the previous years are available on the ADEME website. There is a downward trend, but the number was roughly stable between 2014 and 2019, ranging from 114 gCO_2/km to 112 gCO_2/km.

5.1.3 Routing optimization

We focused on breast cancer patients only, since there are many hospitals capable of treating this pathology. Since we do not have very precise informations on the patients conditions, we chose to optimize for a simple metric: travel distance. The idea was to simulate what would happen if every patient traveled to the closest specialized hospital, while making sure the hospitals capacities were not exceeded. We modeled this problem as an Optimal Transport (OT) task. In the following paragraphs, we first introduce what is OT, and then

explained how we applied it to our problem.

Optimal Transport (OT) is the study of the optimal transportation and allocation of resources. It was introduced in 1781 by the French mathematician Gaspard Monge, [168] who was interested in the problem of the optimal way of redistributing mass. The problem was, given a pile of soil, how can it be transported and reshaped to form an embankment with minimal effort ? During the World War II, the soviet mathematician Leonid Kantorovitch brought major advances in the field [169], by allowing the mass to be split during transportation. A couple of years later, George Dantzig introduced the Simplex Algorithm to solve Linear Programs, including the Kantorovitch Problem. However, solving this Linear Program becomes untractable whenever the dimension is large. In the recent years, an entropic regularization term was added to the OT formulation, allowing to find the optimum in a very fast way [170], using the Sinkhorn-Knopp's algorithm [171].

We now explain more formally how to solve the OT problem with entropic regularization. Consider two distributions α and β, with respectively n and m points x and y, each associated with positive weights a_i and b_j such that $\sum_{i=1}^{n} a_i = \sum_{j=1}^{m} b_i = 1$. The displacement of mass between the two distributions can be described by a set of transport plan, or couplings, defined on Equation (5.1). In this equation, the couplings $U(a, b)$ are the set of transport plan $P \in \mathbb{R}_{+}^{n \times m}$, that satisfies the transportation of mass constraints $P\mathbf{1}_m = a, P^T\mathbf{1}_n = b$. Intuitively, all the mass from the first distribution should be moved to the second distribution. Thus, summing on P column-wise or row-wise should return a and b. We seek to find the transport plan $P \in U(a, b)$ that minimizes the cost Equation (5.2). The first term in this cost is the distance $d(x_i, y_j)^p$ between the two points x_i and y_j. The next term is the Entropic Regularization, weighted by ϵ. The lower ϵ is, the closer we get to the non-regularized OT problem. The minimum solution can be obtained with the Sinkhorn-Knopp's algorithm [171], as explained in [172, 170]. The output of the algorithm is the optimal transport plan σ^*, that moves the input distribution to the output distribution in the most cost effective way.

$$U(a,b) = \{P \in \mathbb{R}_{+}^{n \times m}; P\mathbf{1}_m = a, P^T\mathbf{1}_n = b\} \tag{5.1}$$

$$\min_{P \in U(a,b)} \sum_{i,j} d(x_i, y_j)^p P_{i,j} + \sigma P_{i,j} log(\frac{P_{i,j}}{a_i b_j}) \tag{5.2}$$

In our case, we want to find the optimal way to move patients from their n population locations, to the m hospitals. The distance metric $d(x_i, y_j)$ is the driving distance between the municipality i and the hospital j. The weights a and b correspond to the populations and hospitals capacities respectively. We normalized the populations and capacities so that a and b sum to one. Thus, a_i corresponds to the proportion of patients living in municipality i, and b_j to the proportion of patients that the hospital j can host. The σ^* output matrix contains the overall proportions of patients sent from the municipality i to the hospital j. We multiply each element in this matrix by the total number of patients, and round the result to get the number of patients traveling from the municipality to the hospital.

5.2 Results

5.2.1 Patients travel description

A total of 493,526 patients travels for 12 cancer types were included in the study. The number of distinct population locations was 5,606, and the number of distinct hospitals was 978. The three most frequent pathologies were: malignant melanoma and other malignant skin tumors (n=104,429 stays); malignant breast tumors (n=86,237 stays); and malignant tumors of the digestive organs (n=81,440 stays). The rarest pathologies were malignant tumors of the eye, brain, and other parts of the central nervous system (n=7,904 stays); malignant tumors of mesothelial tissue and soft tissue (n=6,549 stays); and malignant tumors of bone and articular cartilage (n=2,452 stays). We studied the median travel duration, median travel distance, overall distance, number of distinct hospitals and CO_2 emissions by cancer type and hospital oncology specialization. To assess the oncology specialization of the hospitals,

we used the oncology clusters defined in Chapter 2. Hospitals from clusters 1 and 2 are the most oncology specialized hospitals, with all the key services such as cancer surgery, radiotherapy, and chemotherapy. They also have the largest surgeries volumes and are often specialized in even the rarest cancer types. Such hospitals are sparsely located, and often placed in large cities. The hospitals from clusters 3 and 4 are less specialized and are in both large cities and sub-urban areas. The full results are displayed in Table 5.1.

	N stays	Median duration	Median distance	Total distance	N Hospitals	% Hospitals	CO_2 Emissions
Cancer type							
Malignant melanoma and other malignant skin tumors	104,429	21.56	16.18	3,214,375.72	894	91%	360.01
Malignant tumors of the eye, brain and other parts of the central nervous system	7,904	44.39	44.43	616,675.46	327	33%	69.07
Malignant tumors of the lip, oral cavity and pharynx	13,115	29.55	26.35	629,616.37	659	67%	70.52
Malignant tumors of the thyroid and other endocrine glands	9,059	27.57	22.68	405,445.77	564	58%	45.41
Malignant tumors of the digestive organs	81,440	24.31	20.18	3,330,910.43	858	88%	373.06
Malignant tumors of the male genital organs	47,472	24.68	20.66	1,869,128.99	815	83%	209.34
Malignant tumors of the female genital organs	29,501	25.75	21.48	1,249,403.48	799	82%	139.93
Malignant tumors of the respiratory and intrathoracic organs	30,228	31.71	28.69	1,523,374.66	758	78%	170.62
Malignant tumors of bone and articular cartilage	2,452	41.80	39.32	180,105.78	323	33%	20.17
Malignant tumors of the urinary tract	75,140	22.74	17.90	2,565,232.46	803	82%	287.31
Malignant breast tumors	86,237	24.94	20.26	3,290,349.47	810	83%	368.52
Malignant tumors of mesothelial tissue and soft tissue	6,549	33.35	30.04	402,222.65	677	69%	45.05
Hospital Cluster							
Cluster 1: Oncology experts	121,890	33.33	29.73	6,586,967,47	79	8%	737.74
Cluster 2: Oncology experts	38,606	24.46	21.35	1,630,935,96	39	4%	182.66
Cluster 3: Hospitals without radiotherapy	244,493	22.73	18.03	8,377,446,27	451	46%	938.27
Cluster 4: Hospitals without radiotherapy nor chemotherapy	86,245	21.32	16.22	2,634,153,25	348	36%	295.03
Cluster 5: Hospitals with chemotherapy and radiotherapy only	7	15.13	12.17	137.79	2	0%	0.02
Cluster 6: Hospitals with chemotherapy and radiotherapy only	13	30.13	31.55	440.41	3	0%	0.05
Cluster 8: No oncology service	2,272	18.43	11.55	467,60.09	56	6%	5.24

Table 5.1: Patients travel description for each pathology. A total of 493,526 patients travels for 12 cancer types were included in the study. The number of distinct population locations was 5,606, and the number of distinct hospitals was 978. We studied the median travel duration, median travel distance, overall distance, number of distinct hospitals and CO_2 emissions by cancer type and hospital oncology specialization. To assess the oncology specialization of the hospitals, we used the oncology clusters defined in Chapter 2.

For more frequent cancer types, the patients travel remain relatively short, as there are many hospitals with the required specialization. For instance, the shorter travels were for skin tumors patients, with a median distance of 16.18 kilometers and a median duration of 21.56 minutes. Among all the hospitals included, 894 (91.4%) of them performed skin tumor surgeries. However, for the less frequent tumors such as the eye, brain, and other parts of the central nervous system, the patients' travels were longer. Indeed, the median travel duration was 41.8 minutes, and the median distance was 39.32 kilometers. Similarly, the patients' travels were longer when they visit more specialized hospitals, especially cluster 1,

where the median duration is 33.33 km. Patients visiting hospitals from cluster 6 also tend to experience longer travels, with a median duration of 30.13 km. The hospitals within this cluster are hospitals with radiotherapy and chemotherapy activity, but no cancer surgery.

We studied the median driving duration based on the patient municipality of residence. We discretized the median driving duration into 5 bins: < 30 mins; 30-60 mins; 60-90 mins; 90-120 mins; and > 120 mins. On Figure 5.1, map (A) displays the spatial distribution of the median driving duration, in metropolitan France. The municipalities are filled with median driving duration bins. We notice that the duration is lower for patients living in denser municipalities (B). Indeed, the median driving duration for patients living in municipalities with less than 30 inhabitants per km^2 is 50.7 minutes; compared with 16.4 minutes for patients living in municipalities with >200 inhabitants / km^2. We then studied the median travel duration based on patients municipalities density and visited hospital oncology specialization (C). On the alluvium plot (C), we represented patients municipalities grouped by population density on the left, and visited hospitals grouped by oncology cluster on the right. The rectangles sizes are proportional with the number of patients. We colored the alluvium flows based on the median duration. As expected, the driving duration is lowest for patients living in dense municipalities, regardless the hospital they visit. However, for patients living in rural municipalities, the driving duration is higher, especially when they visit hospitals from cluster 1, corresponding to the yellow flow on the plot (C).

5.2.2 Travel burden index

For each patient route, we obtained a travel burden score, expressed as a linear combination between travel duration, travel distance, number of roundabouts and road sinuosity. The weights are the loading scores of the input variables along the first principal component of the PCA analysis. The higher the weight is, the more contribution the input variable has in the component. The loading scores were: 0.57 for duration; 0.55 for distance; 0.32 for number of roundabouts; and 0.52 for road sinuosity. The median travel burden score was 0.069, ranging between 0 and 0.98. We discretized the distribution into 5 quantiles with the

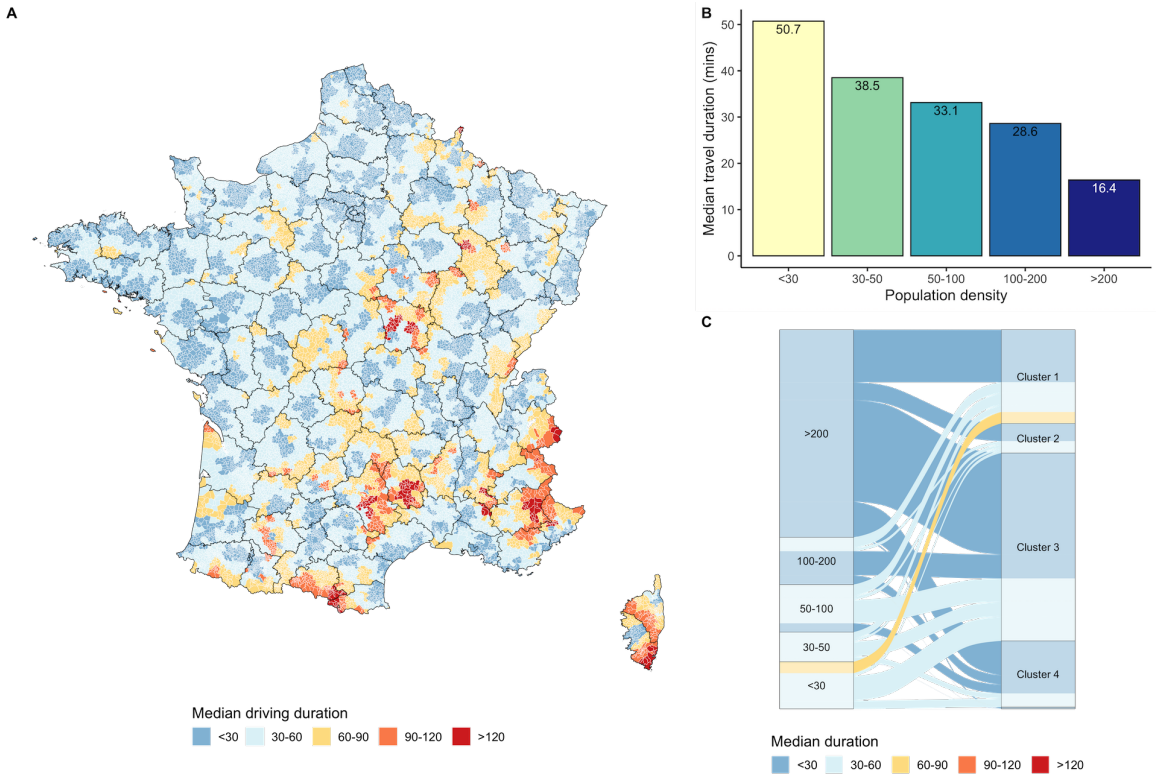


Figure 5.1: Average driving duration for cancer patients in metropolitan France. Map (A) displays the average driving duration by municipalities. The median travel duration is higher for municipalities with lower population densities (B). The median travel duration is especially high for patients from rural areas visiting specialized hospitals (C). Patients living in dense areas do not need to travel far when reaching specialized hospitals (C).

following cuts: 0, 0.04; 0.06 0.08; 0.1; 1. The lower the score, the lower the travel burden is. We discretized the average score into 5 quantiles. For municipalities in the first quantile, the average travel burden score is in the top 20%, meaning that patients travel are shorts and road sinuosity is low. We studied the travel burden score distribution, and compared it to the input variables (Figure 5.2). The score has a strong positive correlation with road sinuosity (0.89); duration (0.85); distance (0.8); and number of roundabouts (0.68). The municipalities median revenue and population densities were lightly negatively correlated, with coefficients of -0.29 and -0.09 respectively (C). We compared the travel duration, distance, sinuosity and score on plots (D) and (E). As expected, we notice that travels in the lowest quantiles have low

duration and distance. However, at a given travel distance, increasing the travel duration will result in a score increase (D). Travel duration exponentially increase with the road sinuosity (E), and high sinuosity values correspond to higher travel burden quantiles.

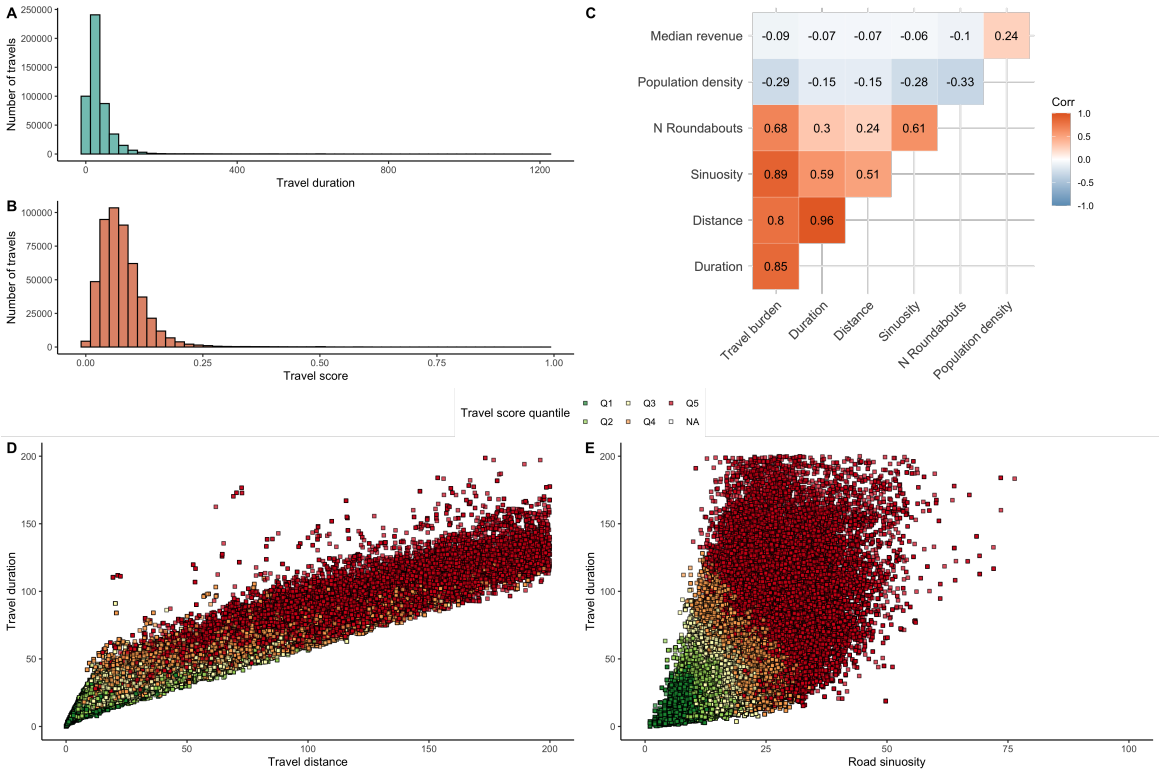


Figure 5.2: Travel burden score distribution per department and region. Comparison between travel duration distribution (A) and travel burden distribution (B). Correlations between travel burden score and other variables (C). Comparison between travel distance, duration, and travel score (D). Comparison between road sinuosity, travel duration and travel score (E).

We now study the spatial distribution of the travel burden score, by displaying the average travel burden score per municipality on a map (Figure 5.3-A). The municipalities are filled by median driving duration, discretized according to the previously stated quantiles. We notice spatial disparities in the distribution. Areas with high travel burden are mostly located in regions like Corse, Pays de la Loire, Occitanie and Provence-Alpes-Cote-d'Azur. In general, municipalities with lower population densities have a higher number of routes with

high travel burden (C). For instance, among all the municipalities with population densities lower than 30 inhabitants per km^2 40.2% are in the worst quantile, compared to only 11.3% for municipalities with 200 or more inhabitants per km^2 (B).

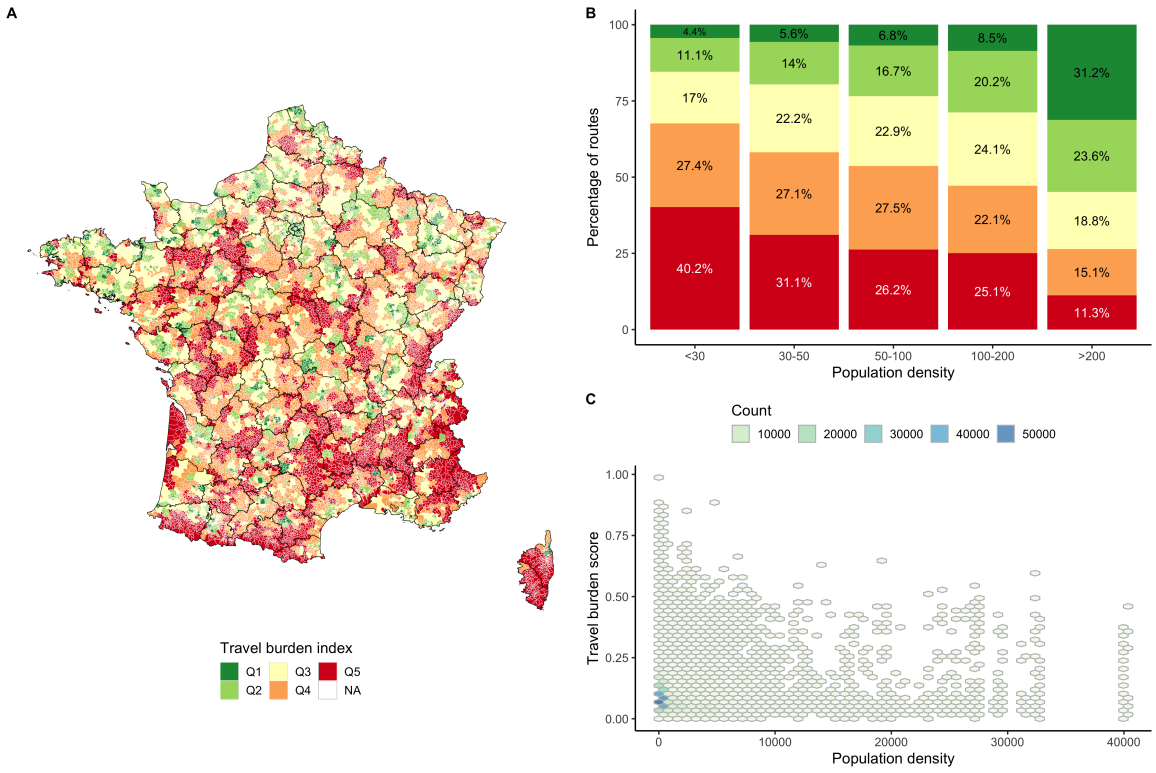


Figure 5.3: Travel burden index in metropolitan France. The travel burden index is a composite score based on route duration, distance, number of roundabouts and sinuosity. The higher the score is, the more tedious the route is. The score distribution is displayed on map (A). The percentage of routes with higher scores increases in lower density areas (B). Figure (C) displays the input variables median values by score quantiles. For instance, the median road sinuosity is much higher when the score is high.

The travel burden index varied from a region to another, especially in the most rural departments. We ranked the regions by median travel burden index in increasing order and obtained the following: Île-de-France (median=0.0477, sd=0.0377); Normandie (median=0.0649, sd=0.0440); Hauts-de-France (median=0.0681, sd=0.0405); Bretagne (median=0.0691, sd=0.0428); Provence-Alpes-Côte d'Azur (median=0.0701, sd=0.0476); Grand Est (median=0.0709,

sd=0.0409); Auvergne-Rhône-Alpes (median=0.0725, sd=0.0492); Bourgogne-Franche-Comté (median=0.0757, sd=0.0466); Nouvelle-Aquitaine (median=0.0773, sd=0.0520); Centre-Val de Loire (median=0.0795, sd=0.0440); Pays de la Loire (median=0.0812, sd=0.0469); Occitanie (median=0.0816, sd=0.0523); Corse (median=0.147 , sd=0.188). The 5 departments which had the lower median travel burden index were Paris, Val-de-Marne, Hauts-de-Seine, Seine-St-Denis, and Rhone. Among these departments, the first 4 are in Ile de France region. The 5 departments with the highest travel burden are from lowest to highest: Aveyron, Corse-du-Sud, Lozère, Ardèche, and Haute-Corse (Figure 5.4).

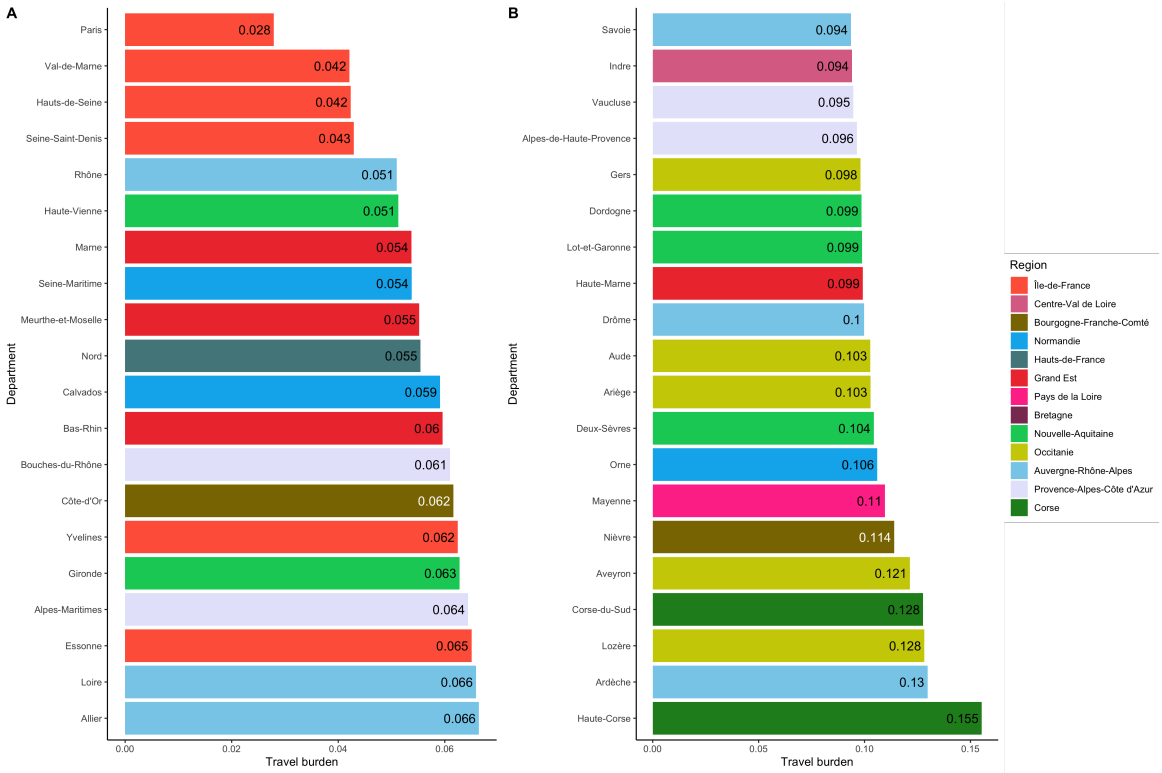


Figure 5.4: Travel burden score distribution per department and region. The 5 departments which had the lower median travel burden index were Paris, Val-de-Marne, Hauts-de-Seine, Seine-St-Denis, and Rhone. Among these departments, the first 4 are in Ile de France region. The 5 departments with the highest travel burden are from lowest to highest: Aveyron, Corse-du-Sud, Lozère, Ardèche, and Haute-Corse

We then focused on a single region, and compared the average travel burden score with

the main roads location, as illustrated on Figure 5.5. We did not show the roads that with were used by less than 5 patients during the year, for clarity. In this region, we recall that the two largest cities are Marseille and Nice, and that the accessibility is the highest along the coastline, where the higher population densities are. The road network is the most developed on the coastline, as well as around cities like Avignon and Gap. The areas that had low accessibility scores have high travel burden scores, which makes sense since the travel burden score was party computed with the travel duration to reach the hospitals. However, we notice that some areas that had decent accessibility scores can have average or high average travel burden scores. This is probably due to the sinuosity of the roads, notably in the Var department, or in the north of Nice city. The roads in these areas are often small, with a lot of turns and roundabouts, increasing the travel tediousness. Overall, the travel burden score is lower for municipalities near the main roads.

5.2.3 Carbon footprint of patients travel

The overall carbon emissions associated with the included travels in this study was 2,159 tons of CO_2. The total emissions per cancer type vary between 373 tons for malignant tumors of the digestive organs, and 20 tons for malignant tumors of bone and articular cartilage. Despite being the cancer type with the most stays, malignant melanoma and skin tumors do not represent the highest carbon footprint (Figure 5.6). Indeed, the 104,429 stays in this pathology are associated with 360 tons of CO_2 emissions; where the 81,440 stays related to malignant tumors of digestive organs are associated with 373 tons of emitted CO_2. The three cancer types with the most stays represent nearly 50% of the overall carbon emissions.

The average CO_2 emissions per travel increased with the rarity of the cancer and the scarcity of hospitals habilitated to treat this disease (Figure 5.7). Indeed, the average CO_2 emissions were the lowest for malignant melanoma and other malignant skin tumors, which had the highest amount of stays (A) and specialized hospitals (B). The rare cancers like bone or eye cancer had the highest average carbon emissions.

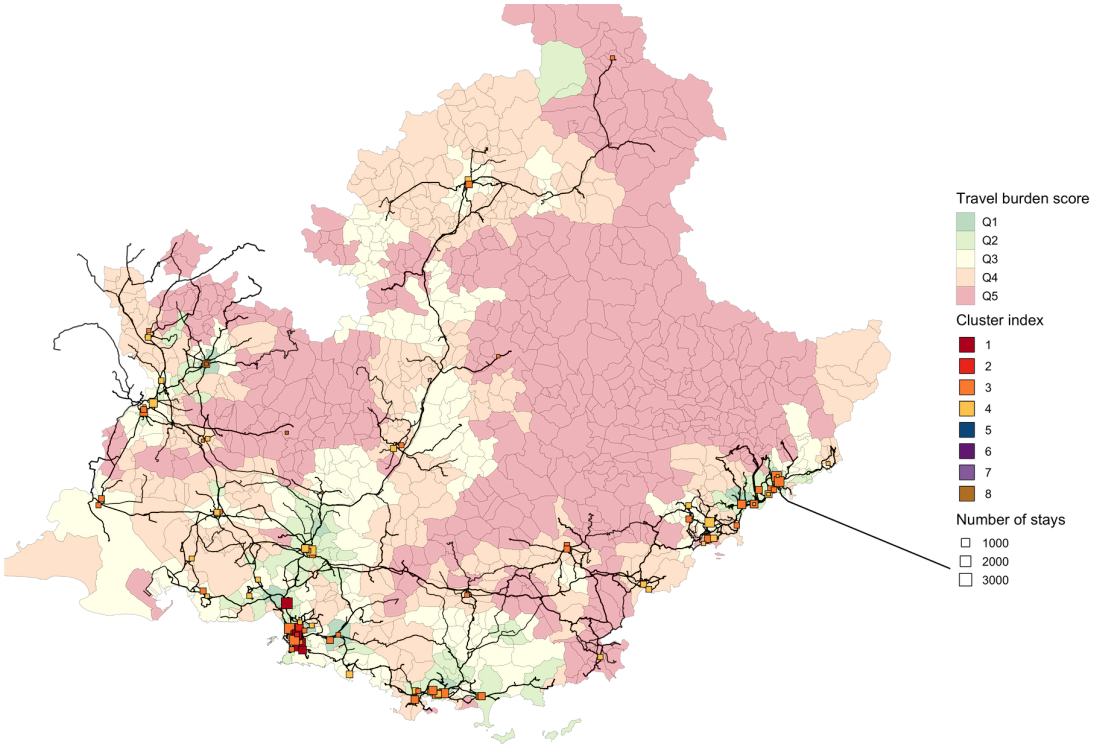


Figure 5.5: Travel burden score in Provence Alpes Cote d'Azur (PACA) region. We compared the average travel burden score with the main roads location. The roads that with were used by less than 5 patients during the year are hidden. The areas that had low accessibility scores have high travel burden scores. However, we notice that some areas that had decent accessibility scores can have average or high average travel burden scores. This is probably due to the sinuosity of the roads, notably in the Var department, or in the north of Nice city. The roads in these areas are often small, with a lot of turns and roundabouts, increasing the travel tediousness.

The average emissions of CO_2 vary depending on the oncology specialization of the visited hospital, and the rurality of the patient municipality of residence. Regardless of the cancer site, the average emissions for patients visiting hospitals from the most specialized cluster was 6.05 kg, versus 3.84 kg and 3.42 kg for hospitals within clusters 3 and 4 (Figure 5.8-A). Similarly, the average emissions for patients living in municipalities with less than 30 inhabitants per km^2 was 8.1 kg, compared with 2.9 kg for municipalities with > 200 inhab-

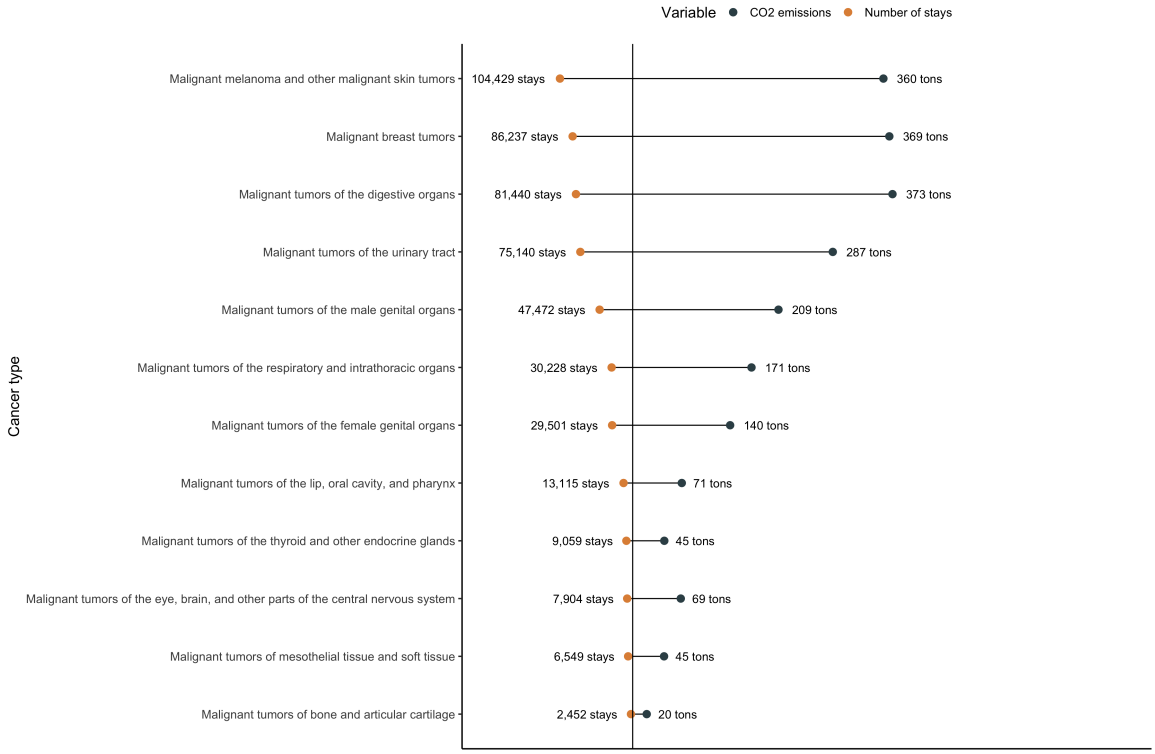


Figure 5.6: Carbon footprint and number of stays by cancer location The total emissions per cancer type vary between 373 tons for malignant tumors of the digestive organs, and 20 tons for malignant tumors of bone and articular cartilage.

itants per km ^{2} (B). Finally, the average emissions were the highest for patients living in the least densely populated areas, and visiting the most specialized hospitals (C). The emissions are 11.1 kg on average for patients living in < 30 inhabitants per km ^{2} and visiting hospitals within cluster 1.

We now focus on a single department and chose Ain, in Auvergne-Rhone-Alpes region. The largest city in Ain is Bourg-en-Bresse, with a population of 41,248 inhabitants in 2018; and a population density of 1,729 inhabitants per km ^{2}. The Ain department is mostly populated with sub-urban and rural municipalities, but it is located near the Rhone department, where urban municipalities are, including Lyon, the third largest city in the country. We analyzed the travels frm patients living in municipalities from the Ain department, and shown which hospitals they visited on Figure 5.9. The alluvium chart on sub-figure (A) displays the

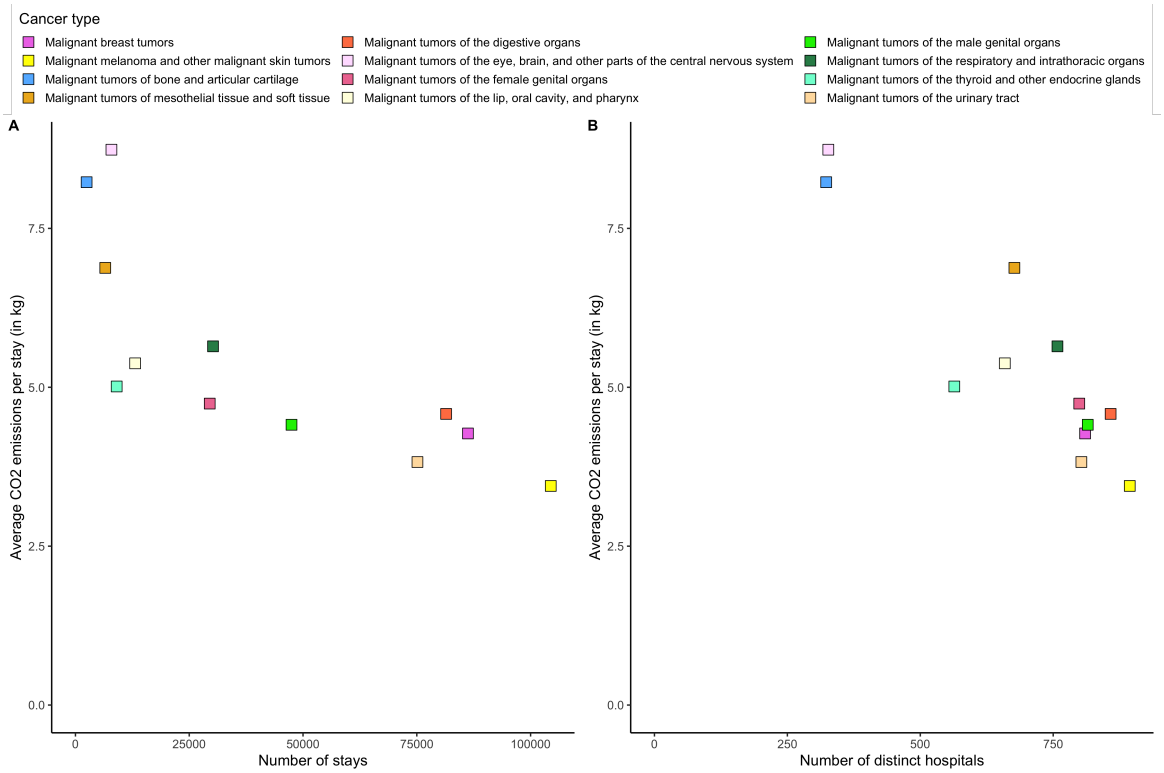


Figure 5.7: Average carbon footprint by cancer location. Comparison between the average CO_2 emissions and the number of stays (A), as well as with the number of habilitated hospitals (B). The average CO_2 emissions per travel increased with the rarity of the cancer and the scarcity of hospitals habilitated to treat this disease.

number of patients routes between municipalities on the left and hospitals on the right. The alluvium flows are sized by number of stays and colored by the stays carbon footprint. The darker flows indicate higher CO_2 emissions. We point that the routes with the more emissions are not necessarily the routes with the most patients. To illustrate this, we show the total CO_2 emissions per visited hospital for patients living in Bourg-en-Bresse (B). Centre-Hospitalier de Fleyriat is the most visited hospitals among patients living in Bourg-en-Bresse, with 150 stays and 4-kilometer distance between the municipality centroid and the hospital. The resulting CO_2 emissions are 72 kg. However, some patients are traveling outside of Bourg-en-Bresse to reach hospitals based in Lyon, which represents at least an 80 km drive. For instance, there were 18 stays in Hospital Lyon Sud, located at 91 km from Bourg-en-

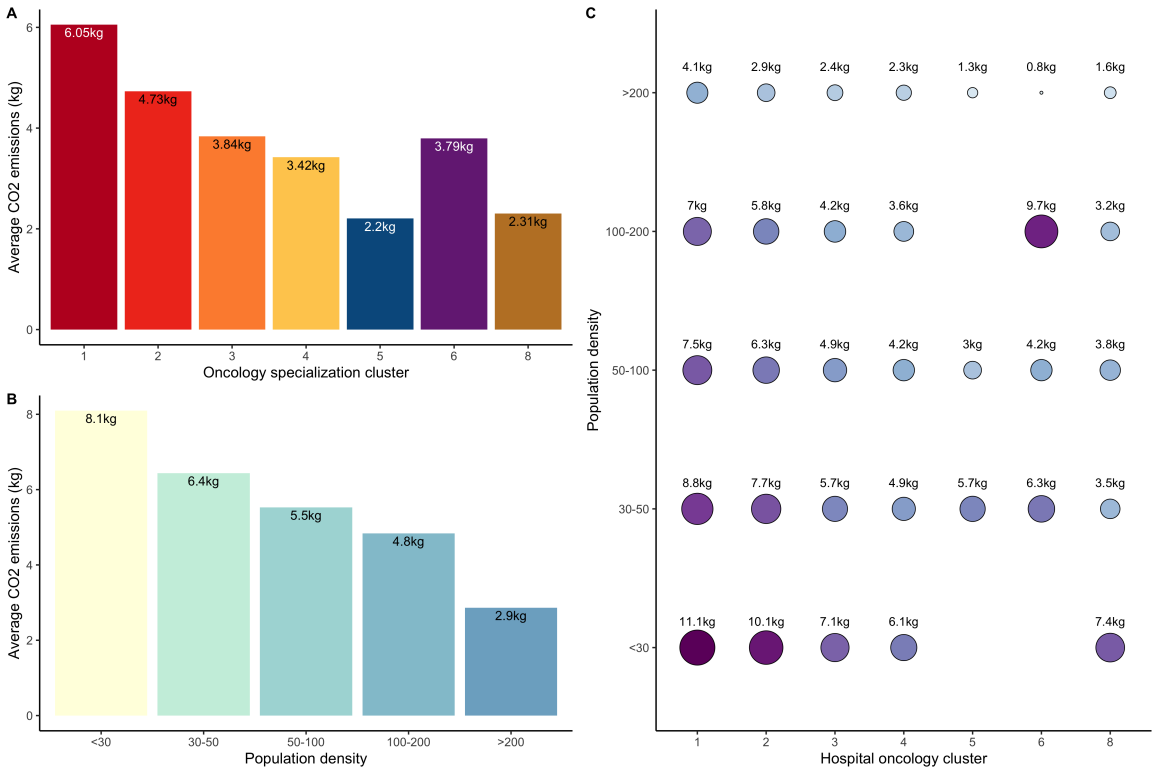


Figure 5.8: Average carbon footprint according to the hospital oncology specialization and municipality population density. Regardless of the cancer site, the average CO_2 emissions are higher for patients visiting the most specialized hospitals (A). Similarly, the emissions are higher for patients living in rural areas (B). Finally, the average emissions are the highest for patients living in rural municipalities and visiting the most specialized hospitals (C).

Bresse. The resulting CO_2 emissions were 184 kg, which is more than twice the emissions of the 150 stays in CH Fleyriat.

5.2.4 Route optimization for cancer patients

We solved the regularized Optimal Transport problem on the matrix of travel durations between population locations and hospitals. We used the ot.bregman.sinkhorn solver from "POT: Python Optimal Transport" [173]. The algorithm performed the patients allocations

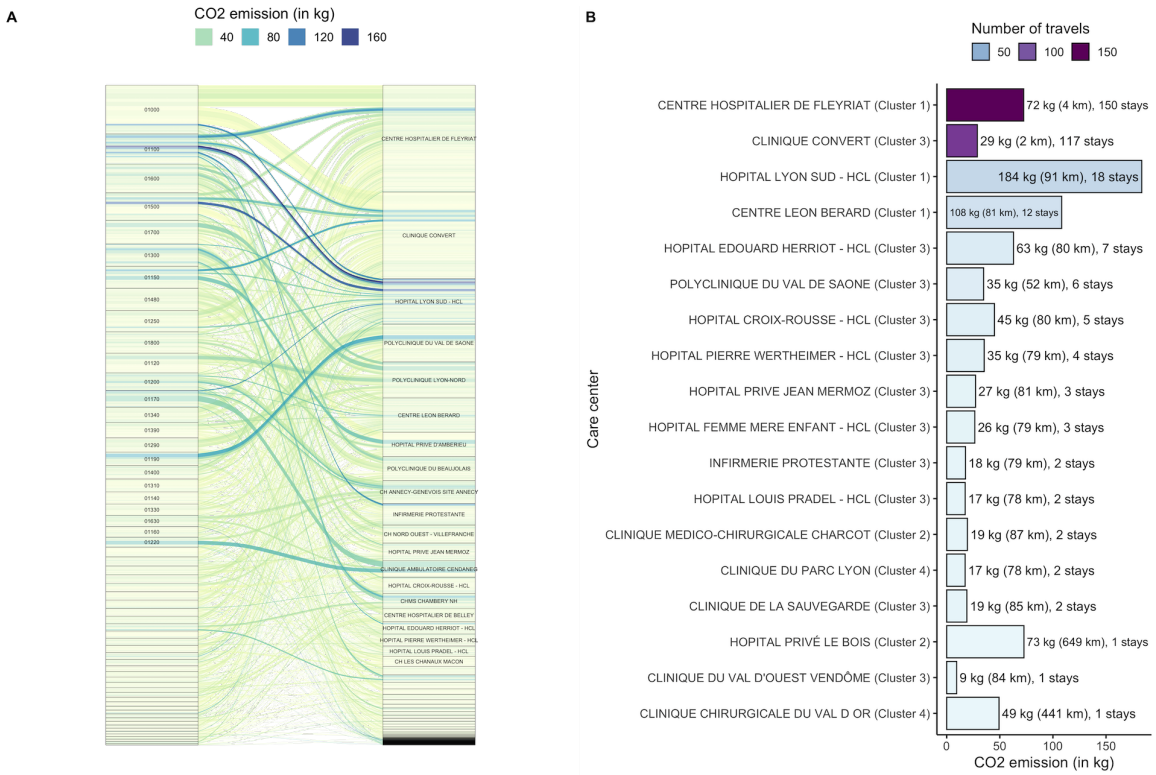


Figure 5.9: CO_2 emissions for cancer patients travels The CO_2 emissions are computed based on the GPS distance between the patient municipality centroid and hospital location. The total emission for a single travel is computed as the product of the average CO_2 emissions per km and the distance. Figure (A) displays the travels between municipalities in Ain department. Municipalities are on the left, hospitals on the right. Flows are sized by number of travels and colored by CO_2 emissions. Figure (B) shows the CO_2 emissions compared with number of stays in Bourg-en-Bresse city (Ain). The CO_2 emissions are higher for the fewer patients who traveled outside of the city to reach more specialized care centers in Lyon.

by minimizing the overall travel duration while respecting the hospitals maximum capacities. The resulting allocations are displayed on Figure 5.10. Map (A) shows the allocations in Provence-Alpes-Cote-d'Azur region. Population locations are displayed as blue triangles, sized by their populations. Hospitals are displayed as red squares, sized by their capacities. Capacities have been defined as the number of patients treated by the hospital within the year. The black lines show the allocations between population locations and patients.

The line width is proportional to the number of patients sent from a population location to an hospital. Since the algorithm minimized the traveled distance, patients tend to visit the closer hospitals. Plot (B) displays the overall traveled distance, and we notice that the optimization process nearly halved the overall distance. The average traveled distance per patient went from 34.5 km to 21.9 km, a 36% decrease. The overall carbon footprint similarly decreased from 293,009 tons of CO_2 to 186,141 tons of CO_2. We compared the travel distance distribution before the optimization (C) and after (D), and notice that very few patients travel further than 250 km with our method.

The alluvial plots on Figure 5.11 display the travels flux between population locations on the right, and hospitals on the left, in the Bouches du Rhone department (PACA region). The boxes are sized by the number of patients living in the municipalities and treated in the hospital. The boxes are sorted by decreasing number of patients. The paths are sized by the number of patients who traveled from the population location to the hospital, and colored by the travel burden quantile. The first alluvial plot on the left (A) displays the routes before the optimization, and the second chart shows the new routing after the OT algorithm (B). Before the optimization process, the travel burden scores were higher for municipalities with lower populations, i.e. located to the bottom of the figure. We also notice that the proportion of patients with more tedious travels is higher for the larger hospitals, especially “Institut Paoli-Calmettes”, which is the most specialized in oncology care within the department. After the optimization algorithm was ran, the proportion of patients with higher travel burden decreased. We also notice that patients are routed more homogeneously. Indeed, patients within the same municipality tend to be sent to the same hospitals.

5.3 Conclusion

We report longer travels for patients living in rural areas. The hospitals specialized in oncology tend to receive patients from more distant population locations. Finally, patients with less frequent cancers are forced to travel further due to the limited number of hospitals that

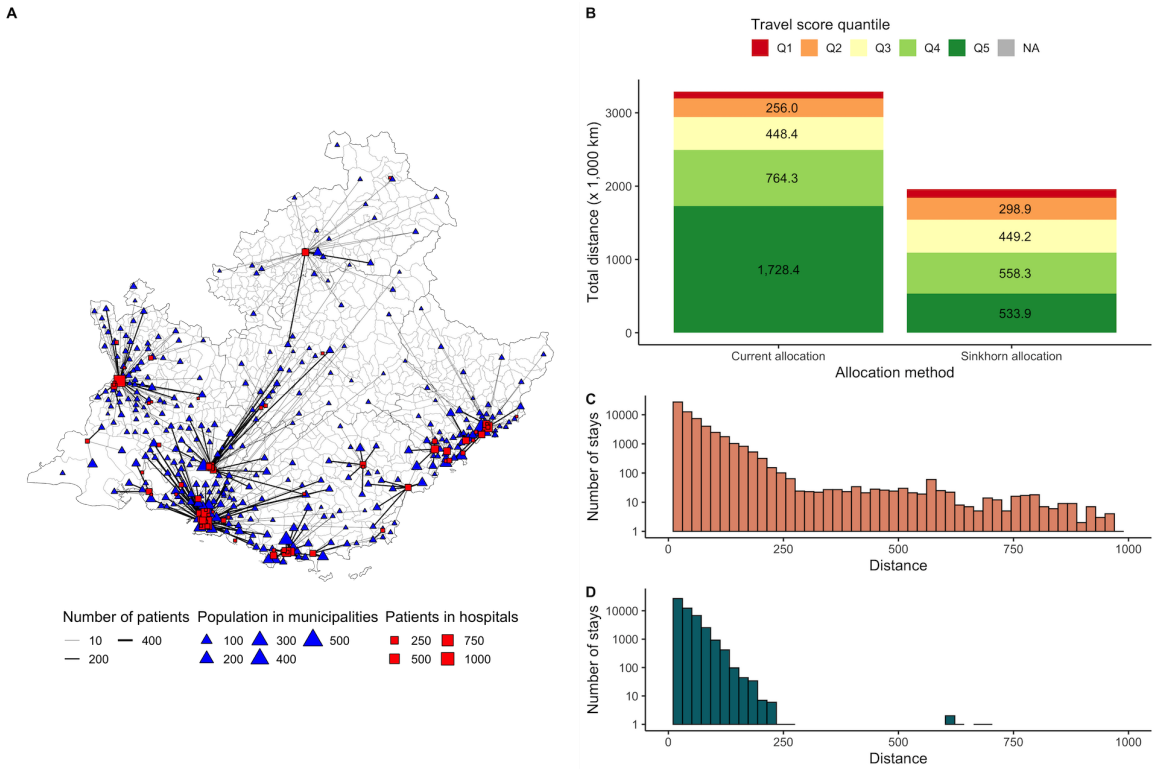


Figure 5.10: Optimization results with the regularized Optimal Transport algorithm. Map (A) shows the allocations in Provence-Alpes-Cote-d'Azur region. Population locations are displayed as blue triangles, sized by their populations. Hospitals are displayed as red squares, sized by their capacities. Plot (B) displays the overall traveled distance, and we notice that the optimization process nearly halved the overall distance. We compared the travel distance distribution before the optimization (C) and after (D), and notice that very few patients travel further than 250 km with our method.

can correctly treat these pathologies. We introduced the travel burden score, a new metric to consider when studying patients travels. This score is proportional to not only distance and duration, but also road sinuosity and number of roundabouts. The last two variables, which are not explicitly captured by distance and duration, could be responsible of more tedious drives for the patients, especially when their health conditions are deprived. In our estimation, the CO_2 emissions are directly proportional to the traveled distance. We did not consider indirect emissions linked to the transportation. Car was the only transportation

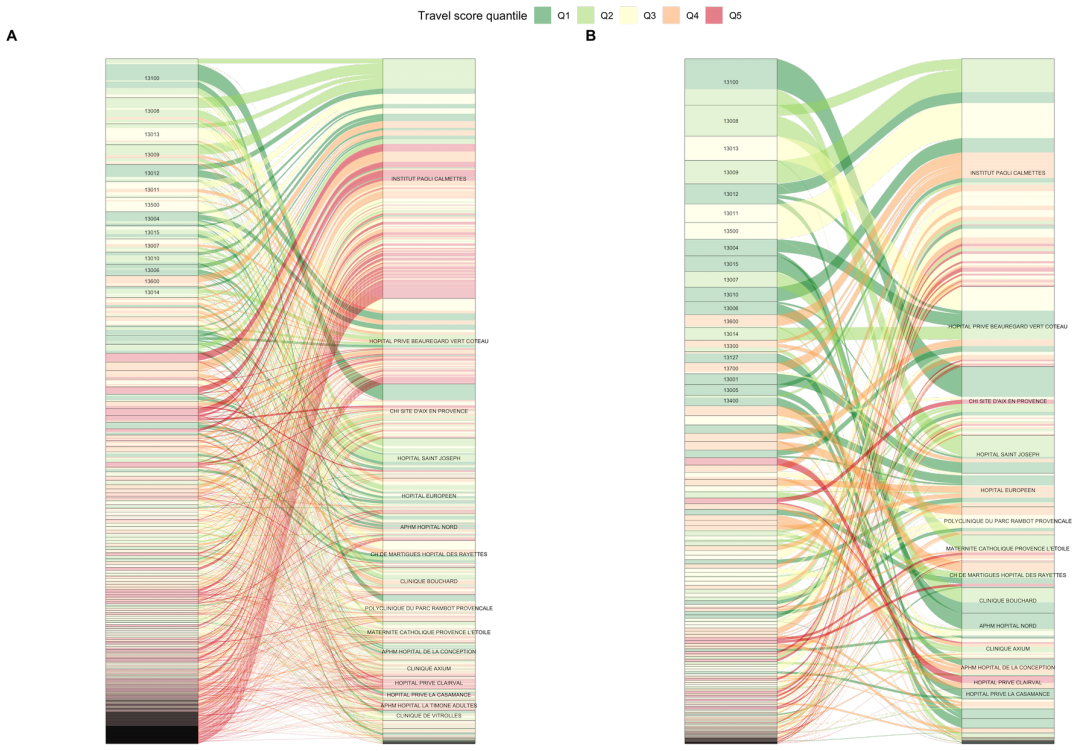


Figure 5.11: Travel flux between in the Bouches du Rhone department (PACA region) before and after optimization. The boxes are sized by the number of patients living in the municipalities and treated in the hospital. The boxes are sorted by decreasing number of patients. The paths are sized by the number of patients who traveled from the population location to the hospital, and colored by the travel burden quantile. The first alluvial plot on the left (A) displays the routes before the optimization, and the second chart shows the new routing after the Optimal Transport (OT) algorithm (B).

mean we used, and we assumed every patient traveled by car, which might over-estimate the CO_2 emissions. More research is needed to include public transportation such as train or subway. The larger share of carbon emissions for cancer surgeries is covered by frequent cancers, that can be treated in many hospitals, like breast cancer for instance. For such pathologies, a rethink of the centralization of care model might be needed. Patients from less dense municipalities could be sent to closer regional hospitals if we make sure the surgeons' expertise is good enough. Partnerships with larger and more specialized hospitals

could be created to spread the more up to date knowledge outside the urban hospitals. However, this will be more complicated for rare cancers, where expertise is scarce and concentrated in the larger hospitals. On a carbon footprint perspective, we believe the lower number of concerned patients makes it less of a priority. Finally, we simulated the case where every patient would travel to the closest hospital, provided we do not exceed the hospitals maximum capacities. We showed that the average driving distance and CO_2 emissions were reduced by 36%. While these results are promising, only minimizing the traveled distance is not sufficient to route the patients to the optimal hospital. More factors should be taken into account, such as hospital specialization, quality of care, and detailed patients characteristics. By comparing the number of patients by hospital before and after the optimization algorithm, we noticed that the largest and most specialized hospitals received less patients than before. These hospitals are often saturated, and lowering the number of patients they receive could benefit them as well as the patients treated there. These new vacancies could also be filled by patients with more complicated cases or rare cancers that require a specific expertise that not every hospital have. We are now interested in the global effects of our optimization algorithm. A tradeoff should be found between travel distance and patient-hospital affinity. The case we presented where the patients traveled to the nearest hospital is the most optimistic situation, and despite this the driving distance and associated CO_2 emissions are “only” reduced by 36%. Only considered surgery stays were considered here, thus telemedicine will not be usable to reduce the footprint. The only lever to reduce the associated carbon footprint is the average CO_2 consumption of the driving vehicles, which will probably drop with the democratization of the electric cars. To sum up, the results of the travel analysis for cancer patients in metropolitan France concur with the effects of centralization of care observed in the literature, where patients living in rural areas tend to experience longer drives, that are also more tedious.

Chapter 6

Transparency in healthcare

This chapter will be part of a research article, currently being written.

6.1 Methods

With these evidences of healthcare information needs, we developed Healthcare Network, a web application that lists every hospital in France, and displays key statistics on them. The application is directed to either health professionals or patients. Health professionals might use it to gain insights about specific hospitals, and look for the best place to send their patients when they lack expertise. Patients could learn more about the hospital they have been sent to, check the care quality or surgery volume.

To create the Healthcare-Network web application, we centralized the several datasets into databases. We then built the backend of the application with Python and Flask framework, while the frontend was coded with HTML and CSS from the Bootstrap library. We used two databases: a relational database (MySQL) and a no relational database (Mongo DB). In Mongo DB, we stored the datasets to draw the interactive maps in the application. We used the geojson format, which works well with Mongo DB. All the other datasets are stored in the relational database. There are roughly 40 tables in the relational database. The most used tables are statistics on the hospitals and on the municipalities. We chose to use the

Flask framework for its simplicity and the high number of users. The framework has a simple core to quickly build a working web application with little code. It can be improved with extensions that add new features. With such framework, we built several API routes that render HTML and CSS templates to the users. The interactive maps were built with the Folium library, and the other figures with Plotly.

6.2 Results

In this section, we will describe the most important features of the web application, illustrated with screen captures from the website.

The landing page of the application is a minimalist search bar, as illustrated on Figure 6.1. The search feature lets the user look for hospitals by name, category or location. In this first version, the search algorithm is fairly basic and matches the hospitals that contains the text query in either its name category or location. No weighting or additional rule was added. The search feature could be improved later on, by using dedicated and search focused databases like the very popular Elastic Search.

Figure 6.1: Healthcare-Network: homepage. A minimalist page with a search bar allowing to find hospitals based on their name, category, or location.

We now provide a search example, and Figure 6.2 shows the results of the query “chu”,

corresponding to the Centre Hospitalier Regional / Universitaire (CHR/U) hospital category. 153 results were returned, displayed as a list on the left and on a map on the right side of the screen. The list shows basic informations about the retrieved hospitals, including their names, locations, categories and type of care provided, like Medecine, Chirurgie, Obstetrique (MCO) for instance. The map is interactive, and lets the user visualize the spatial distribution of the retrieved hospitals. On the map, the hospitals are displayed as icons, colored by hospital category. This query illustrates an additional feature of the search bar: show the spatial distribution of the hospitals in France by category.

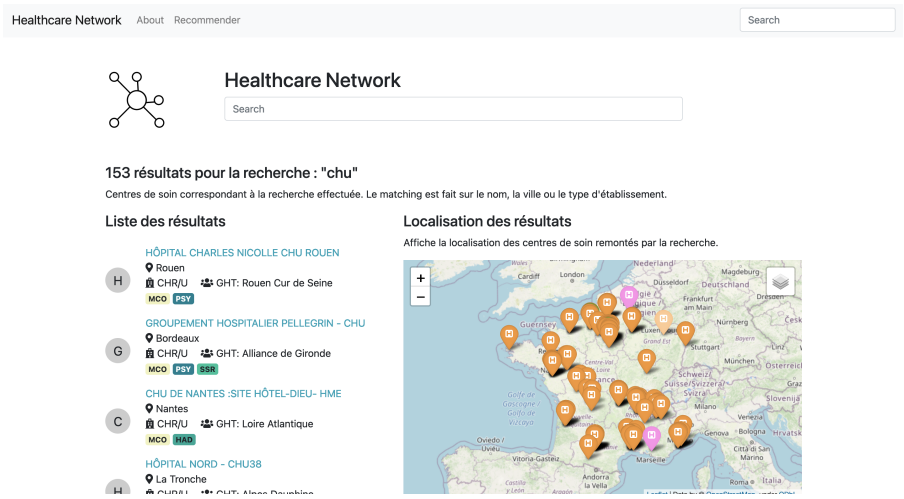


Figure 6.2: Healthcare-Network: search results. The list of retrieved hospitals and their details is displayed, as their position on a map. This query shows all the CHR/U hospitals in metropolitan France.

Each hospital has its individual web page, as illustrated on Figure 6.3 with the CH de Coulommiers. The web page shows basic informations about the hospital, with name, location and category displayed first. A navigation pane also shows the hospitals from the same Groupement Hospitalier de Territoire (GHT) and legal entity. Hospitals within the same GHT share their information system. This grouping was introduced in 2016 for the public hospitals only. They aimed at facilitating the communication between these facilities, and make it easier to transfer patients from one hospital to another if needed. Hospitals from the

same legal entity are governed by the same administration, but spread among multiple geographical sites. Hospitals in large cities such as Paris, Marseille or Lyon have most of their largest hospitals belonging to the same legal entity. For instance in Paris, the AP-HP legal entity gathers 39 hospitals spread across the Ile-de-France region. The hospital location is shown on an interactive map, where the user can zoom in and out, and add more indicators, including:

- The municipalities populations and median salary. These indicators are a way to gain insight about the hospital neighborhood, and neighboring demand. To display these indicators, we color the municipality according to the indicator value.

- Patients provenance. We display the number of patients who visited this hospital per municipalities within a year. Through this, it is easy to evaluate how influent and important an hospital is, based on how many patients it is draining from further population locations. Usually, small local hospitals tend to receive patients from their immediate neighborhood; where large hospitals specialized in oncology like Institut Curie or Institut Gustave Roussy will treat patients from many different regions.

- Other hospitals from the same GHT. We display on the map the other hospitals that share the same information system. With this information, we can evaluate how close this hospital is to other hospitals where it would be easy to transfer patients if the desired pathology is not treated in this hospital.

- Other hospitals that shared patients with this hospital within a year. We call this 'co-occurrences', and a higher number shows that two hospitals seems to work closely together. For instance, one hospital might handle the cancer surgery and send their patients to another hospital for radiotherapy. Identifying hospitals that frequently exchange patients is a good way to find alternative hospitals for certain pathologies. There is also a high chance that these two hospitals communicate frequently with each other, making it easier to send patients and keep track of what happened in their pathways.

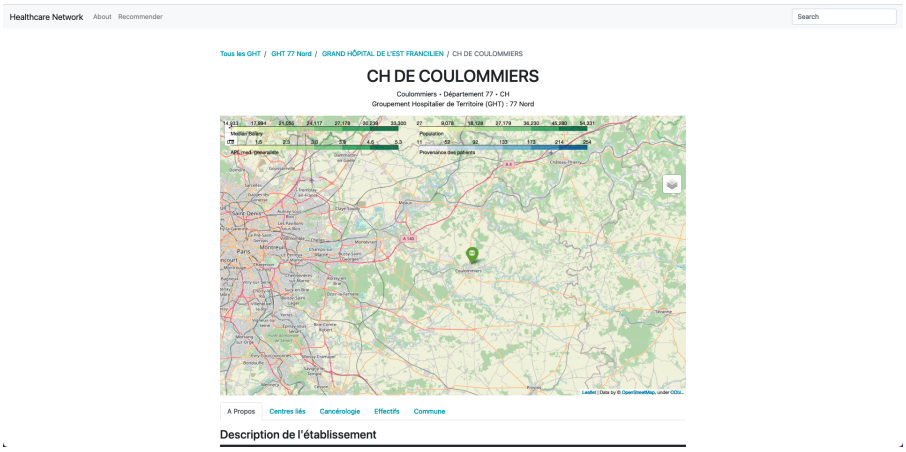


Figure 6.3: Healthcare-Network: example of an hospital page, Centre Hospitalier (CH) de Coulommiers. The web page shows basic informations about the hospital, with name, location and category displayed first. A navigation pane also shows the hospital GHT and legal entity. The hospital location is shown on an interactive map.

The Figure 6.4 is an example of the interactive map mentioned earlier. Here, we colored the municipalities by the number of patients who visited CH de Coulommiers. We also displayed hospitals from the same legal entity in green. Finally, we show hospitals that exchanged patients with CH de Coulommiers as blue links. From this map, we observe that CH de Coulommiers mostly attract patients from the neighboring municipalities. The hospitals with the most co-occurrences are either from the same GHT, or located in Paris. This might mean that patients with complications were sent to larger hospitals in Paris.

Below this interactive map, we display the list of services and number of MCO stays and beds for hospitals. For instance, on Figure 6.5, we displayed this information for the Institut Curie Paris hospital. The list of services allows to quickly evaluate the hospital ability to treat cancer patients. The number of stays and number of beds lets the users evaluate the hospital size, and how saturated it is. We see that Institut Curie has all the main services besides psychiatry and palliative care. From the activity statistics, we see that this hospital does not do any obstetric activity.

Next, we display a section focused on oncology activity, we show statistics related to

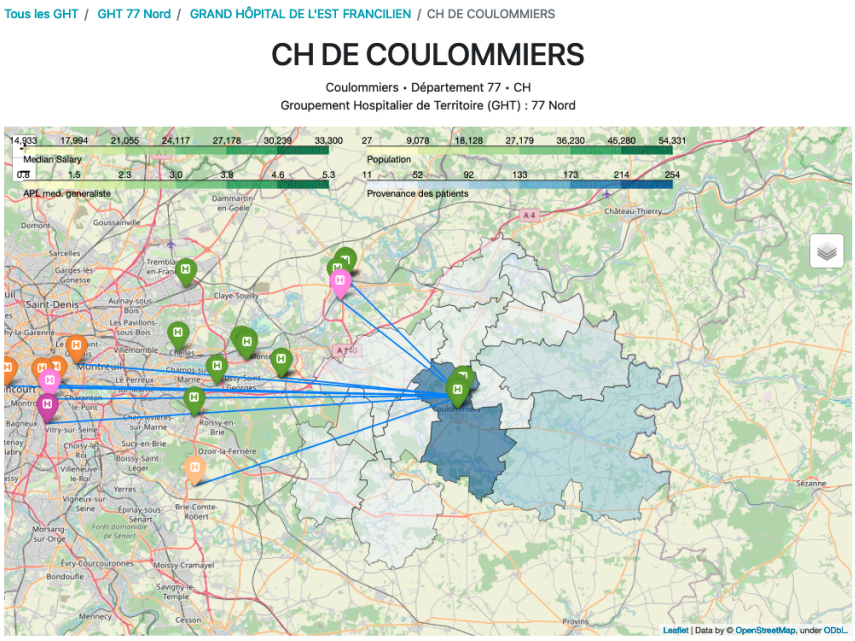
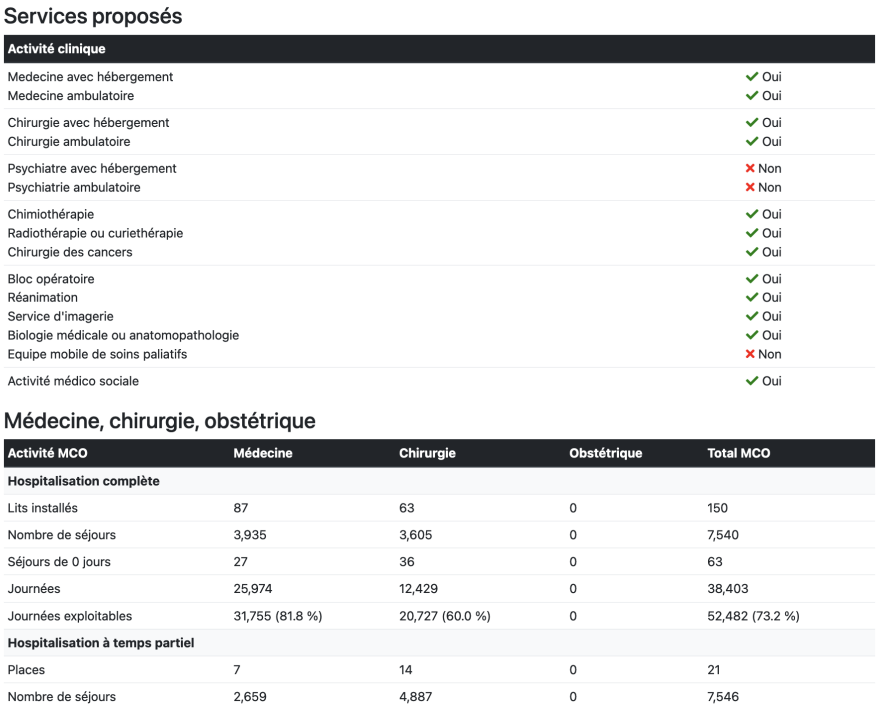


Figure 6.4: Healthcare-Network: example of an hospital page, Centre Hospitalier (CH) de Coulommiers. We filled the municipalities by the number of patients who visited CH de Coulommiers. We also displayed hospitals from the same legal entity in green. Finally, we show hospitals that exchanged patients with CH de Coulommiers as blue links.

cancer care, as illustrated on Figure 6.6. This section lists the key oncology services like radiotherapy, cancer surgery and chemotherapy authorization. The number of cancer related stays, the number of radiotherapy stays as well as the number of beds dedicated to oncology are shown.

Finally, we show the number of patients stays by cancer organs treated in the hospital. We chose a radar chart visualization, as displayed on Figure 6.7. Three series are shown on this plot. First, in green, we show the statistics of the current hospital. Then, in purple, we display the median number of patients treated for every hospital in the same category, while the orange curve shows the overall median. Showing these three variables allows the users to compare the current hospital with hospitals within the same category as well as broader comparison to the overall median. In this case, the numbers are from Institut Curie hospital. The radar chart shows that this hospital treats more patients than the other hospitals from



Services proposés

Activité clinique	
Medecine avec hébergement	✓ Oui
Medecine ambulatoire	✓ Oui
Chirurgie avec hébergement	✓ Oui
Chirurgie ambulatoire	✓ Oui
Psychiatre avec hébergement	✗ Non
Psychiatrie ambulatoire	✗ Non
Chimiothérapie	✓ Oui
Radiothérapie ou curiethérapie	✓ Oui
Chirurgie des cancers	✓ Oui
Bloc opératoire	✓ Oui
Réanimation	✓ Oui
Service d'imagerie	✓ Oui
Biologie médicale ou anatomopathologie	✓ Oui
Equipe mobile de soins paliatifs	✗ Non
Activité médico sociale	✓ Oui

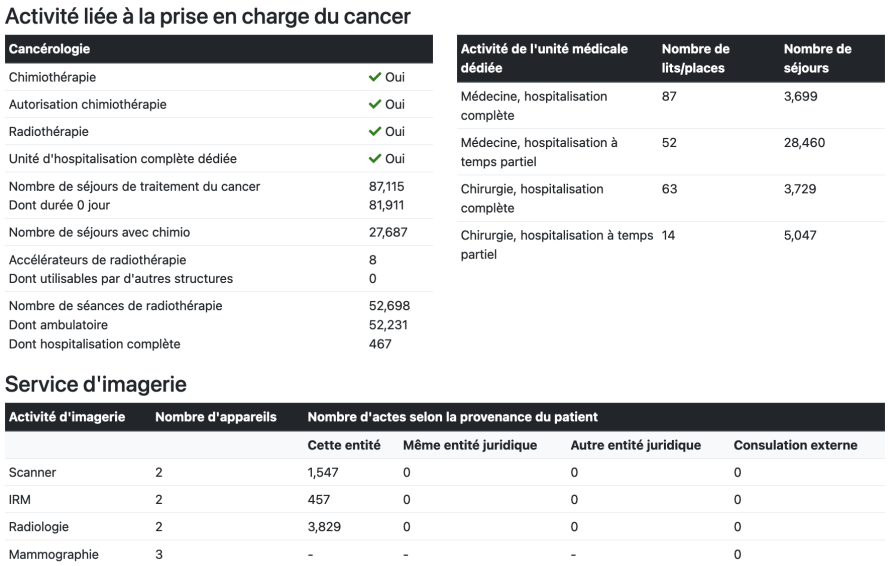
Médecine, chirurgie, obstétrique

Activité MCO	Médecine	Chirurgie	Obstétrique	Total MCO
Hospitalisation complète				
Lits installés	87	63	0	150
Nombre de séjours	3,935	3,605	0	7,540
Séjours de 0 jours	27	36	0	63
Journées	25,974	12,429	0	38,403
Journées exploitables	31,755 (81.8 %)	20,727 (60.0 %)	0	52,482 (73.2 %)
Hospitalisation à temps partiel				
Places	7	14	0	21
Nombre de séjours	2,659	4,887	0	7,546

Figure 6.5: Healthcare-Network: description of health services offered, and statistics on MCO activity for Institut Curie Paris hospital. The list of services allows to quickly evaluate the hospital ability to treat cancer patients. The number of stays and number of beds lets the users evaluate the hospital size, and how saturated it is.

the same category, especially for eye cancer, where Institut Curie is among the only hospitals with expertise on this rare cancer.

To gain insights about the hospital neighborhood, we show socio-demographic statistics on the municipality where the hospital is located. In the case of Institut Curie, the municipality is the 5th arrondissement of Paris. Among the numbers displayed, we have the municipality size, population, median salary and poverty rate. We also count the number of health professionals by occupation within the hospital department, and within the hospital. This gives information on the link between the hospital and the town medicine which is very important for patient care.



Activité liée à la prise en charge du cancer

Cancérologie	
Chimiothérapie	✔ Oui
Autorisation chimiothérapie	✔ Oui
Radiothérapie	✔ Oui
Unité d'hospitalisation complète dédiée	✔ Oui
Nombre de séjours de traitement du cancer Dont durée 0 jour	87,115 81,911
Nombre de séjours avec chimio	27,687
Accélérateurs de radiothérapie Dont utilisables par d'autres structures	8 0
Nombre de séances de radiothérapie Dont ambulatoire Dont hospitalisation complète	52,698 52,231 467

Activité de l'unité médicale dédiée	Nombre de lits/places	Nombre de séjours
Médecine, hospitalisation complète	87	3,699
Médecine, hospitalisation à temps partiel	52	28,460
Chirurgie, hospitalisation complète	63	3,729
Chirurgie, hospitalisation à temps partiel	14	5,047

Service d'imagerie

Activité d'imagerie	Nombre d'appareils	Nombre d'actes selon la provenance du patient			
		Cette entité	Même entité juridique	Autre entité juridique	Consulation externe
Scanner	2	1,547	0	0	0
IRM	2	457	0	0	0
Radiologie	2	3,829	0	0	0
Mammographie	3	-	-	-	0

Figure 6.6: Healthcare-Network: description of oncology activity for Institut Curie Paris hospital.

6.3 Conclusion

Our Healthcare-Network web application could benefit both patients and healthcare professionals. It might incentive health professionals to find a closer hospital from the patient location, lowering both travel burden as well as CO_2 emissions. Also, patients will be able to double check where they have been sent to, which could reduce dissatisfaction and health disparities. In Healthcare-Network, hospitals statistics are displayed in the most transparent way possible. This often involves showing plain numbers to patients or health professionals, which might be confusing and unclear for some. More work might be needed to make sure the web application is intuitive and brings useful information to everyone. Moreover, while we developed the web application with the help of medical experts, we did not gather patients' feedback yet. The app design and features are thus likely to change in the future. While our web application brings more health information to the patients, we should be cautious about undesired effects of this approach. Indeed, research showed that the increase in the amount of available medical information resulted in some difficul-

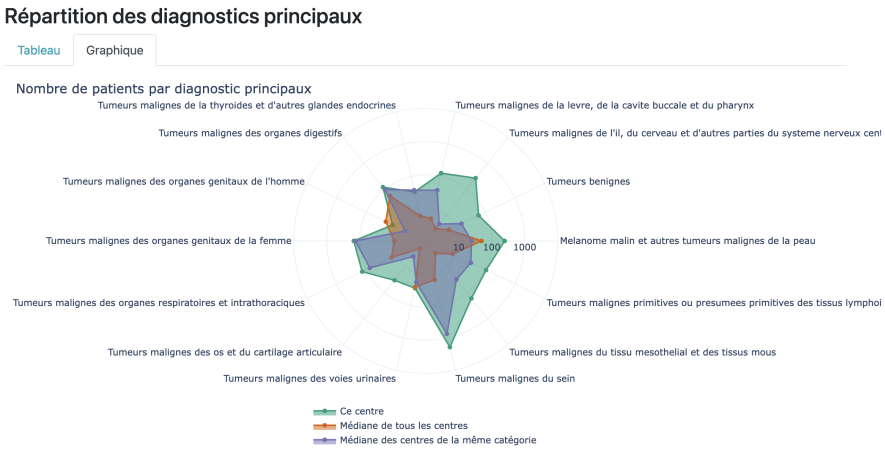
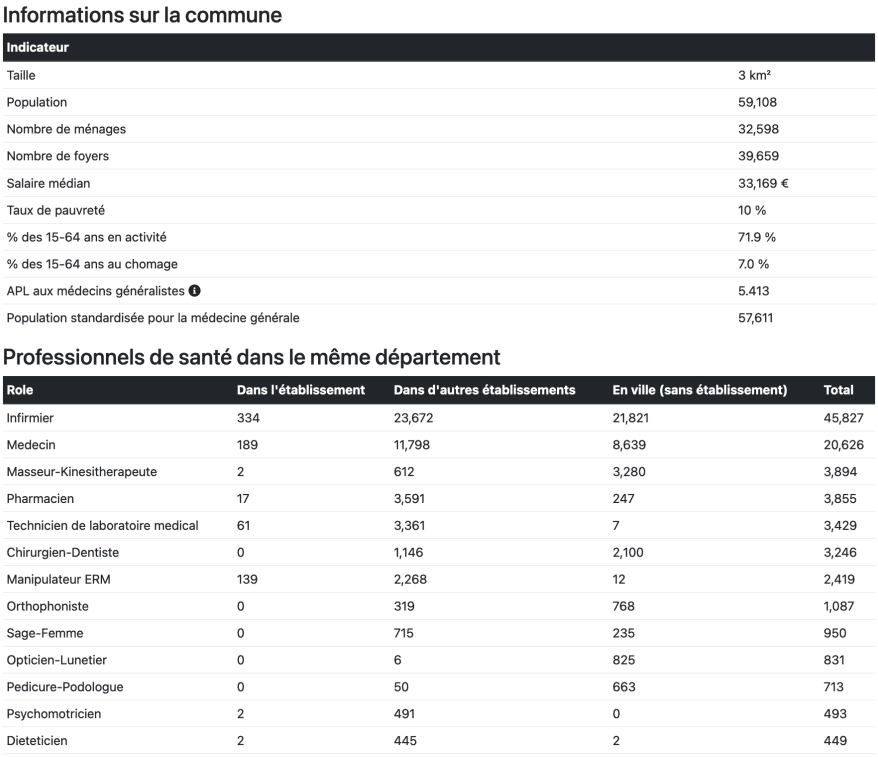


Figure 6.7: Healthcare-Network: number of patients per cancer related diagnosis for Institut Curie Paris hospital. Comparison with the median statistics from hospitals within the same category (CLCC) and overall median.

ties for patients when search-ing for suitable doctors [174, 175]. This gap opened the need for patient-doctor matchmaking, in which patients can find the right doctors based on several criteria [176]. Recommender systems are a typical way to solve such problems. They have been integrated into online retailers, streaming services, and social networks to facilitate users' item selection process. Recently, these systems have been widely applied to the healthcare domain to help both end-users and medical professionals in making more efficient and accurate health related decisions [177]. Recommender systems have also been used to provide personalized doctor recommendations based on emotions and preferences of users about doctors through their ratings and reviews [178]. Although the current literature has shown many benefits of Health Recommender System (HRS) to improve their health conditions, there still exist some gaps regarding developing and evaluating HRS that need to be bridged [177], especially during their evaluation [179]. Uncertainty in HRS links to potential risks and imprecise predictions since user preferences are not al-ways captured well. For these reasons, we chose not to embed a hospital recommender system in the web application yet, since the information we have on the patients are not detailed enough. Showing the health facilities nearby the patients' location as well as some key statistics is



Informations sur la commune

Indicateur	
Taille	3 km²
Population	59,108
Nombre de ménages	32,598
Nombre de foyers	39,659
Salaire médian	33,169 €
Taux de pauvreté	10 %
% des 15-64 ans en activité	71.9 %
% des 15-64 ans au chomage	7.0 %
APL aux médecins généralistes	5.413
Population standardisée pour la médecine générale	57,611

Professionnels de santé dans le même département

Role	Dans l'établissement	Dans d'autres établissements	En ville (sans établissement)	Total
Infirmier	334	23,672	21,821	45,827
Medecin	189	11,798	8,639	20,626
Masseur-Kinesitherapeute	2	612	3,280	3,894
Pharmacien	17	3,591	247	3,855
Technicien de laboratoire medical	61	3,361	7	3,429
Chirurgien-Dentiste	0	1,146	2,100	3,246
Manipulateur ERM	139	2,268	12	2,419
Orthophoniste	0	319	768	1,087
Sage-Femme	0	715	235	950
Opticien-Lunetier	0	6	825	831
Pedicure-Podologue	0	50	663	713
Psychomotricien	2	491	0	493
Dieteticien	2	445	2	449

Figure 6.8: Healthcare-Network: statistics on the municipality where Institut Curie Paris is located (Paris 75105). Population, median salary and accessibility to primary care are displayed to qualify the hospital neighborhood. Health professionals within the department are also listed to illustrate the health supply available around the hospital.

a good first step to assist patients and health professionals during the hospital selection. Moreover, while mass media communications can be an important source of health information, research found social disparities in health knowledge that may be related to media use. Indeed, cancer-related health communications seems to be patterned by race, ethnicity, language, and social class. The benefits of health information are not equally distributed across socially distinct groups in the United States [180]. A lower likelihood of cancer information seeking was also observed among those with lower education levels, lower income and greater ages [110]. Therefore, we must make sure our tool can be accessed by most of the patients, and efforts must be made in addressing these social disparities. Improving healthcare by making it more connected will not be sufficient as it will not benefit a non-

insignificant part of the population. The full benefits of a more connected and transparent healthcare will show when the more deprived populations can access such tools.

Chapter 7

Sinkhorn Matrix factorization with Capacity constraints (SiMCa)

This chapter is part of a research article currently released as a preprint on arXiv.

7.1 Related work

Hospital and practitioner recommendation has already been studied in the literature (see e.g. the survey [177]). However, to the best of our knowledge, no existing method incorporates hospital capacity constraints in the algorithm training. This tends to refer many users to the same hospital, potentially saturating it and degrading the overall care quality.

Matrix factorization [132] is among the most popular collaborative filtering recommendation algorithms. Matrix factorization characterizes every user i and item j by high-dimensional embeddings u_i, v_j, and predict the user-item affinity by the inner product $\langle u_i, v_j \rangle$. This method has already been applied for patient/doctor recommendation [178, 176]. However, regular matrix factorization is usually applied to simple recommendation problems, such as movie recommendation: as already explained before, recommending locations brings new challenges and requires a different approach [134].

Geographical influence has been integrated in the matrix factorization framework to rec-

ommend locations or points of interest (POIs) [181]: moreover, the learning algorithm can be adapted by adding a capacity term in the loss function [133].

The Monge-Kantorovitch formulation of the classical OT problem can be rephrased as a linear program that can be computationally slow and unstable in high dimension [170]: this problem is often approximated by adding an entropy regularization term, and easily solved by Sinkhorn-Knopp's algorithm [170]. Another important advantage of this regularization is that the solution of the OT problem becomes differentiable with respect to the parameters, which explains why this step is integrated in many learning algorithms [182, 183, 184].

Most relevant for the present paper is the work from Dupuy, Galichon and Sun [185]. In this study, the authors address the inverse optimal transport problem, that is, given vectors of characteristics $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^{d'}$ and the joint distribution of the optimal matching, the problem of recovering the affinity function of the form $\phi(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T \mathbf{A} \mathbf{Y}$, namely to estimate matrix $\mathbf{A}$. The authors are in the setting where they observe pairs of embeddings $(\mathbf{X}_t, \mathbf{Y}_t)$ together with the optimal *regularized matching* π^* – that is the solution to problem (7.2) hereafter – and build an estimator of $\mathbf{A}$ with low-rank constraints, the objective being to isolate important characteristics that carry the most important weight in the matching procedure between x and y. We stress the fact that the setting is different in our study: we only observe in our case the embeddings $\mathbf{U}$ of the users and a distance matrix $\mathbf{D}$, function ϕ is known as well as the *pure matching* σ^* – that is the solution of the linear assignment problem (7.1) hereafter, which differs from π^* – and the aim is to infer item embeddings $\mathbf{V}$. In other words, we do not seek to reconstruct the affinity matrix, but for the learning of items' positions in the user's embeddings space, these positions acting as reference points, upon which prediction of future allocations can be made. Another difference is that the number of items is typically very small compared to the number of users, which justifies that the items are considered static: we also incorporate *capacity constraints* on the allocation problem.

7.2 Problem definition

A model for latent and geographical affinity

The setting of the problem is as follows. Consider n *users* $x_1, \dots, x_n$ embedded in a latent space $\mathcal{X}$ identified to $\mathbb{R}^d$, with embeddings given by $\mathbf{U}_1, \dots, \mathbf{U}_n$. Also consider m *items* $y_1, \dots, y_m$ embedded in $\mathcal{X}$ with embeddings $\mathbf{V}_1, \dots, \mathbf{V}_m$, with $m \leq n$. To each user x_i we assign a single item y_j, according to an *affinity matrix* $\mathbf{M} \in \mathbb{R}^{n \times m}$ given by

$$\mathbf{M}_{i,j} := \Phi(\mathbf{U}_i, \mathbf{V}_j, \mathbf{D}_{i,j}),$$

where $\mathbf{D} \in \mathbb{R}^{n \times m}$ is known and may be thought of e.g. as a geographical distance matrix between users and items in the underlying euclidean space, say $\mathbb{R}^2$ (we stress the fact that this space is *not* the embedding space $\mathcal{X}$). We will denote $\mathbf{M} = \Phi(\mathbf{U}, \mathbf{V}, \mathbf{D})$ in the sequel.

We also work under the following constraints: each item $y_j, j \in [m]$ can be assigned to at most $\mathbf{C}_j$ users. Where $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_m)$ is *capacity* vector. The *total capacity* is defined by

$$s(\mathbf{C}) := \sum_{j \in [m]} \mathbf{C}_j,$$

and we will assume $s(\mathbf{C}) = n$. We define

$$\Sigma(n, m, \mathbf{C}) := \left\{ \sigma \in \{0,1\}^{n \times m},\ \sigma \mathbf{1}_m = \mathbf{1}_n, \sigma^T \mathbf{1}_n = \mathbf{C} \right\}.$$

In the sequel, σ will denote both the assignment and its corresponding matrix representation. The optimal assignment σ^* is given by

$$\sigma^*(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{C}) := \underset{\sigma \in \Sigma(n,m,\mathbf{C})}{\arg\max}\ \mathrm{Tr}\left(\sigma^T \mathbf{M}\right), \tag{7.1}$$

Note that problem (7.1) is an instance of the *Linear Assignment Problem (LAP)*.

Goal

Assume that we are given the user embeddings $\mathbf{U}$, the distance matrix $\mathbf{D}$, the capacities $\mathbf{C}$ and the optimal assignment $\sigma^* \in \Sigma(n, m, \mathbf{C})$. The goal is to learn the item embeddings $\mathbf{V}$.

Loss metrics, regularization and relaxation

We will evaluate the performance of a proposed estimate $\widehat{\mathbf{V}}$ of $\mathbf{V}$ through the assignment $\widehat{\pi}$ obtained with $\widehat{\mathbf{V}}$. To compare $\widehat{\pi}$ with σ^*, we use the usual *cross entropy loss* defined by

$$H(\sigma^*, \widehat{\pi}) := -\sum_{i \in [n]} \log \widehat{\pi}_{i,\sigma^*(i)} = -\mathrm{Tr}\left((\sigma^*)^T (\log \widehat{\pi})\right).$$

As stated before, from a learning perspective, a main issue is that the solution to problem (7.1) is not differentiable w.r.t. $\mathbf{V}$, the variable of interest. This issue is solved by a relaxation/regularization procedure [170]:

- since the objective function is linear, we first consider the classical relaxation of (7.1) on the polytope of the convex hull of $\Sigma(n, m, \mathbf{C})$, namely on

$$\Pi(n, m, \mathbf{C}) := \left\{\pi \in [0,1]^{n \times m},\ \pi \mathbf{1}_m = \mathbf{1}_n, \pi^T \mathbf{1}_n = \mathbf{C}\right\}.$$

- moreover, we regularize the objective function in order to perform (automatic) differentiation: this is made possible by the classical entropy regularization in optimal transport.

For a small regularization parameter $\varepsilon > 0$, the problem then becomes

$$\pi^*_\varepsilon(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{C}) := \underset{\pi \in \Pi(n,m,\mathbf{C})}{\arg\max} \left[\mathrm{Tr}\left(\pi^T \mathbf{M}\right) + \varepsilon H(\pi)\right], \tag{7.2}$$

where

$$H(\pi) := -\sum_{1 \le i,j \le n} \pi_{i,j}(\log \pi_{i,j} - 1). \tag{7.3}$$

It is known in the literature [170] that the solution π_ε^* to the convex optimization problem (7.2) can be easily computed with Sinkhorn-Knopp's algorithm, and has the following form:

$$(\pi_\varepsilon^*)_{i,j} = a_i \exp\left(\frac{1}{\varepsilon}\mathbf{M}_{i,j}\right) b_j, \tag{7.4}$$

where a and b are vectors of $\mathbb{R}_+^n$ and $\mathbb{R}_+^m$. Note that we are back to our initial problem (7.1) when $\varepsilon = 0$.

SiMCa Algorithm

With this new formulation (7.2), we are now able to design an optimization scheme for our learning problem. In our setting the users embeddings $\mathbf{U}$, the distance matrix $\mathbf{D}$ and the capacities $\mathbf{C}$ are known, only the items embeddings $\mathbf{V}$ are learned. The overall procedure is summarized in Algorithm 1. Given the current estimate $\mathbf{V}_t$ at iteration t, we compute the solution $\pi_\varepsilon^*(\mathbf{V}_t)$ to problem (7.2), which in turn is used to compute the gradient in $\mathbf{V}_t$ of the following loss

$$\text{loss}(\mathbf{V}_t) := H\left(\sigma^*, \pi_\varepsilon^*(\mathbf{V}_t)\right) \tag{7.5}$$

to update our estimate of $\mathbf{V}$ through a gradient step. The gradient in $\mathbf{V}$ has actually a simple analytical expression:

Lemma 1. *We have*

$$\nabla_{\mathbf{V}}\text{loss}(\mathbf{V}) = \frac{1}{\varepsilon}\sum_{1\le i,j\le n}(\pi_\varepsilon^*(\mathbf{V}) - \sigma^*)_{i,j}\nabla_{\mathbf{V}}\mathbf{M}_{i,j}\,. \tag{7.6}$$

Proof. A very similar expression for the gradient is derived for the maximum likelihood in [185]. We straightforwardly adapt their derivation to the cross entropy loss (7.5). Let us denote

$$V_\varepsilon(\mathbf{M}) = \max_{\pi\in\Pi(n,m,\mathbf{C})}\left[\text{Tr}\left(\pi^T\mathbf{M}\right) + \varepsilon H(\pi)\right] \tag{7.7}$$

the optimal value of the regularized OT problem (7.2). As well-known in the OT literature,

see Proposition 9.2 of [172], its gradient with respect to the affinity matrix M is given by the optimal coupling

$$\frac{\partial}{\partial \mathbf{M}_{i,j}} V_\varepsilon(\mathbf{M}) = (\pi_\varepsilon^*)_{i,j}. \tag{7.8}$$

Our cross-entropy loss (7.5) is directly related to the optimal value $V_\varepsilon(\mathbf{M})$:

$$\begin{aligned}
\text{loss} = H\left(\sigma^*, \pi_\varepsilon^*\right) &= -\sum_{i,j} \sigma^*_{i,j} \ln(\pi^*_\varepsilon)_{i,j} \\
&\overset{1}{=} -\sum_{i,j} \sigma^*_{i,j} (\tfrac{1}{\varepsilon}\mathbf{M}_{i,j} + \ln a_i + \ln b_j) \\
&\overset{2}{=} -\sum_{i,j} \sigma^*_{i,j} \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - \sum_{i,j} (\pi^*_\varepsilon)_{i,j} (\ln a_i + \ln b_j) \\
&\overset{3}{=} -\sum_{i,j} \sigma^*_{i,j} \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - \sum_{i,j} (\pi^*_\varepsilon)_{i,j} (\ln(\pi^*_\varepsilon)_{i,j} - \tfrac{1}{\varepsilon}\mathbf{M}_{i,j}) \\
&\overset{4}{=} -\sum_{i,j} \sigma^*_{i,j} \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - s(\mathbf{C}) \\
&\quad - \sum_{i,j} (\pi^*_\varepsilon)_{i,j} (\ln(\pi^*_\varepsilon)_{i,j} - 1) + \sum_{i,j} (\pi^*_\varepsilon)_{i,j} \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} \\
&\overset{5}{=} -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[\mathrm{Tr}(\pi_\varepsilon^{*T}\mathbf{M}) + \varepsilon H(\pi^*_\varepsilon) - \mathrm{Tr}(\sigma^{*T}\mathbf{M})] \\
&\overset{6}{=} -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[V_\varepsilon(\mathbf{M}) - \mathrm{Tr}(\sigma^{*T}\mathbf{M})].
\end{aligned}$$

The first and third equalities follow from (7.4), the second and fourth from $\sigma^*, \pi^*_\varepsilon \in \Pi(n, m, \mathbf{C})$, the fifth from the definition (7.3) of $H(\pi)$ and the sixth from the definition (7.7) of $V_\varepsilon(\mathbf{M})$. Then differentiating with respect to $\mathbf{V}$ leads to (7.6) by the chain rule and (7.8). □

The performance of our method is guaranteed by the following:

Lemma 2. *Assume that $v \mapsto \Phi(u, v, d)$ is linear. Then the loss function* (7.5) *is convex in $\mathbf{V}$ and the output of SiMCa Algorithm (Algo. 1) converges to*

$$\arg\min_{\mathbf{V}} H\left(\sigma^*, \pi^*_\varepsilon(\mathbf{V})\right).$$

Algorithm 1 SiMCa

Input: $\mathbf{U}, \mathbf{D}, \mathbf{C}, \sigma^*$
For $t = 1$ to T:

1. Compute the affinity matrix $\mathbf{M}_{t-1} = \Phi(\mathbf{U}, \mathbf{V}_{t-1}, \mathbf{D})$.
2. Compute the solution to the optimization problem (7.2):

$$\pi_\varepsilon^*(\mathbf{V}_{t-1}) := \underset{\pi \in \Pi(n,m,\mathbf{C})}{\arg\max} \left[\mathrm{Tr}\left(\pi^T \mathbf{M}_{t-1}\right) + \varepsilon H(\pi) \right].$$

3. Compute the gradient $\nabla \mathrm{loss}(\mathbf{V}_{t-1})$ with equation (7.6).
4. Perform a gradient step $\mathbf{V}_t = \mathbf{V}_{t-1} - \eta \nabla \mathrm{loss}(\mathbf{V}_{t-1})$.

return $\mathbf{V}_T$

Proof. The proof of Lemma 1 shows that

$$\mathrm{loss}(V) = -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[V_\varepsilon(\mathbf{M}) - \mathrm{Tr}(\sigma^{*T}\mathbf{M})].$$

Since $V \mapsto \Phi(\mathbf{U}, V, \mathbf{D})$ is linear, $V \mapsto V_\varepsilon(\mathbf{M})$, as defined in (7.7) is convex as a maximum of convex functions. By assumption, $V \mapsto \mathrm{Tr}(\sigma^{*T}\mathbf{M})$ is linear, thus $V \mapsto \mathrm{loss}(V)$ is convex. □

7.3 Illustration for the hospital recommendation problem

We now describe an illustration of our method for the hospital recommendation problem. Since very few open datasets are available for this problem, we trained our algorithm on synthetic data.

Dataset generation

The dataset is generated as follows:

- **Features in the embedding (latent) space:** we sample $n + m$ points from a Gaussian

mixture model with k clusters. We set these points as either users ($\mathbf{U}_i$) or items ($\mathbf{V}_i$), and considered that each cluster must contain at least one item: we are thus left with n users and m items, spread between k clusters. Users and items in the same cluster are considered similar. We then normalized both users and items features, so that all embeddings $\mathbf{U}_i$ and $\mathbf{V}_j$ lie on the unit sphere. Note that the users and items sampling is done independently of items capacities.

- **Distance in the underlying euclidean space:** to sample the distance matrix $\mathbf{D}$ between users and items, we sample all the positions randomly on a circle, and computed the great-circle distance (i.e. spherical distance) between every users i and items j. We finally normalize the distance matrix by its overall mean.

- **Capacities** we sampled m values from a Dirichlet Distribution, corresponding to the probabilities that users are assigned to the m items. We converted these probabilities into capacities $\mathbf{C}_j$ by multiplying them with the number of users n. We then added some extra spots to each item.

Affinity matrix

In our case, the affinity matrix $\mathbf{M} = \Phi(\mathbf{U}, \mathbf{V}, \mathbf{D})$ is defined as follows:

$$\mathbf{M}_{i,j} = \Phi(\mathbf{U}_i, \mathbf{V}_j, \mathbf{D}_{i,j}) = (1-\alpha)\mathbf{U}_i^T\mathbf{V}_j - \alpha\mathbf{D}_{i,j}. \tag{7.9}$$

The α coefficient measures the trade-off between affinity and proximity.

We then solve the Linear Assignment Problem (7.1) to compute the pure matching σ^*.

Noise

Noise is added to the original dataset in two different ways. The first method is to modify the allocations of random users in σ^*, the noise ratio being defined as the percentage of

modified allocations[1]. The second method consists in perturbating $\mathbf{U}$ as follows:

$$\widetilde{\mathbf{U}} := \sqrt{1-\rho^2}\mathbf{U} + \rho\mathbf{Z},$$

where $\mathbf{Z}$ is a matrix with i.i.d. standard Gaussian entries, and ρ is the noise ratio.

Learning the embeddings

Given $\mathbf{U}$, $\mathbf{D}$, $\mathbf{C}$, σ^*, α and ε, we compute an estimate $\widehat{\mathbf{V}}$ of the item embeddings with SiMCa Algorithm (Algorithm 1). Comparing $\widehat{\mathbf{V}}$ with $\mathbf{V}$ gives a first measure of the training performance.

Recovering the pure matching

Then, using $\widetilde{\mathbf{U}}$ (the noisy version of $\mathbf{U}$), $\widehat{\mathbf{V}}$ (the estimated $\mathbf{V}$), $\mathbf{D}$, α and ε, we compute the solution $\widehat{\pi}^*_\varepsilon$ to problem (7.2). Solving the LAP on matrix $\widehat{\pi}^*_\varepsilon$, we compute a pure matching $\widehat{\sigma}^*$, which we can next compare to the original σ^*, giving a second measure of the training performance.

7.4 Results

Parameters

We generated a toy dataset with the following parameters: $n = 1000$ users; $m = 3$ items; $d = 2$ latent features; $k = 3$ clusters; $\alpha = 0.3$. The items capacities were 257, 417 and 356. Figure 7.1 shows the generated users and items in both the embeddings (latent) space and their underlying euclidean space.

[1] to make sure that the capacities constraints on the items still hold, we must swap *pairs of users*: for a given allocation to modify, we pick another user randomly and swap their allocations.

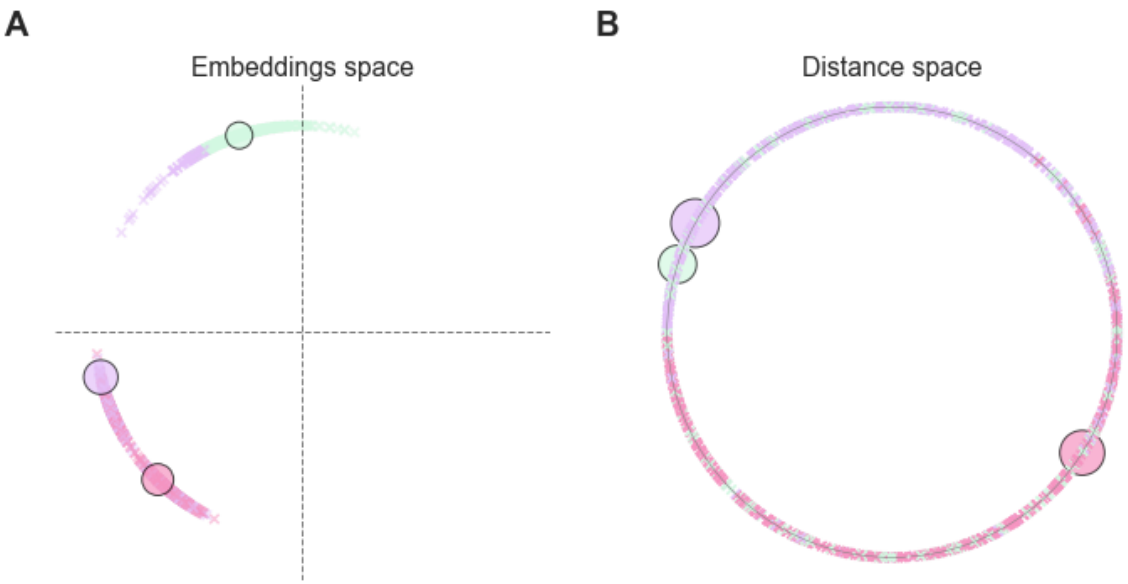


Figure 7.1: Generated dataset. Users are displayed as crosses and items as circles, sized proportionally to their capacities. Users are colored accordingly to the item they have been allocated to. Plot (A) displays users and items in their shared embeddings (latent) space; where plot (B) displays them in their underlying euclidean space.

Training results

We trained our model with $\varepsilon = 0.1$ entropy regularization and 10 iterations in Sinkhorn-Knopp's algorithm to output $\widehat{\mathbf{V}}$. As mentioned earlier we compute a solution to the LAP on matrix $\widehat{\pi}^*_\varepsilon$ to output the estimated allocation $\widehat{\sigma}^*$. The model was trained with Adam optimizer, with a 0.01 learning rate and 400 epochs. For the measures of performance, we used the F1 score, for measuring how well the allocations are reproduced, and the mean euclidean distance between learned embeddings $\widehat{\mathbf{V}}$ and the ground truth $\mathbf{V}$. The training results are displayed on Figure 7.2.

Influence of entropy regularization

We investigate the influence of the entropy regularization parameter ε on the model performance. We let ε vary between 0.05 and 2, with the same dataset and the same model

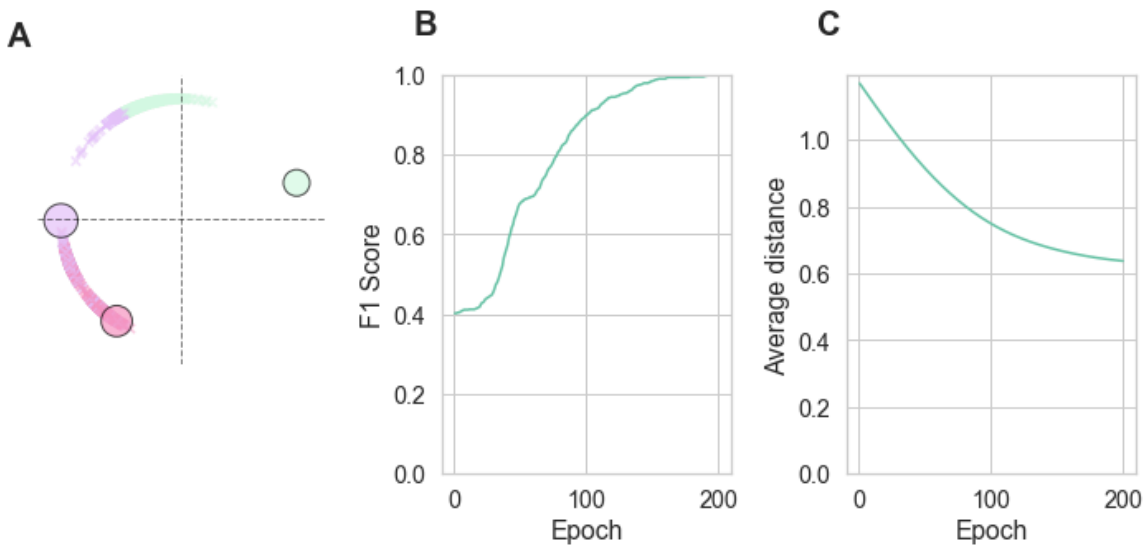


Figure 7.2: Training results. We can see that the model achieves good performances to learn the item embeddings (A), and recovers the allocation with a close to 1 F1 score (B). The average distance between the learned embeddings and ground truth decreases during training (C).

parameters[2]. We ran 5 training for every value of ε. As shown on Figure 7.3, the training performance worsens when ε increases.

Influence of noise

We study the influence of noise, either by swapping allocations or adding Gaussian noise to the used embeddings, as described in the previous section. Unsurprisingly, as shown in Figure 7.4, the training performance is decreasing with the noise ratio.

Learning both users and items embeddings

We also studied the case where the users' embeddings are not known, and must be learned jointly with the items' embeddings from the observed allocations. In this case, we initialized the items' and users' embeddings similarly. We managed to retrieve the observed allocation

[2] due to numerical instability, the algorithm could not train properly above $\varepsilon = 0.05$.

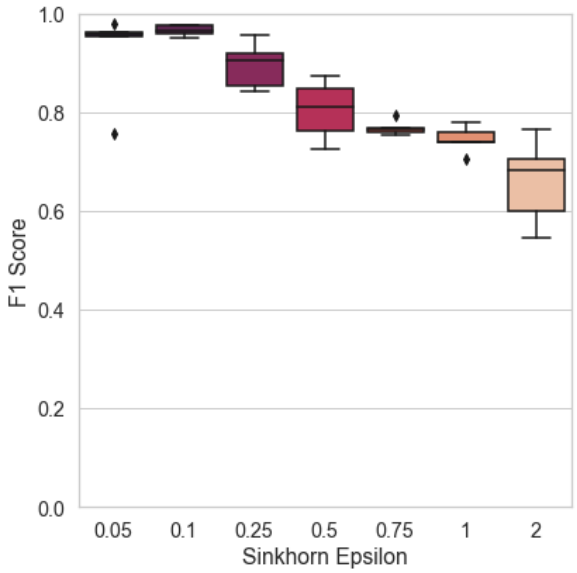


Figure 7.3: Performance as a function of ε. Increasing ε leads to lower F1 scores.

as illustrated on Figure 7.5. However, the average distance between learned items embeddings and ground truth does not decrease during training, meaning that the model learned its own interpretation of the users' and items' representations to satisfy the observed mapping.

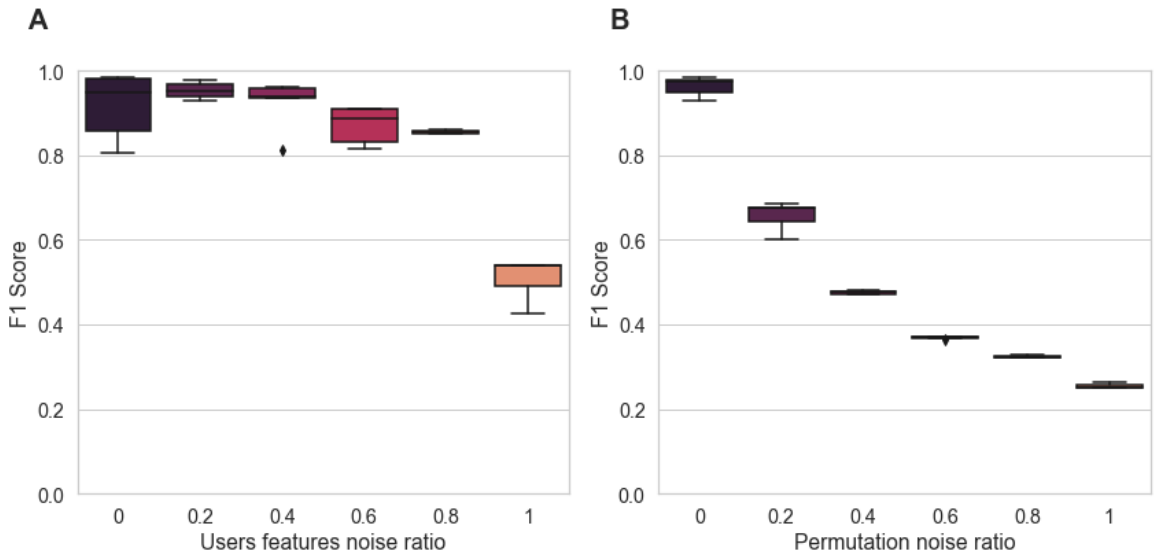


Figure 7.4: Influence of adding Gaussian noise (A) and swapping allocations (B) on training performance. F1 score decreases as noise increases.

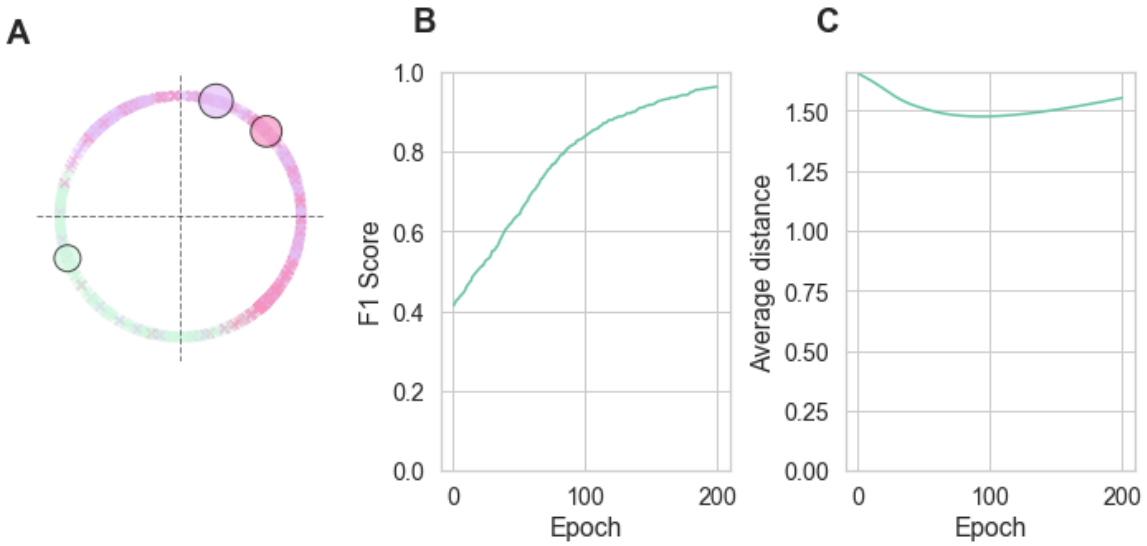


Figure 7.5: Learning both users and items embeddings simultaneously. The learned embeddings are shown on (A). The model retrieves the observed allocation (B). However, the average distance between learned items embeddings and ground truth does not decrease during training (C).

Conclusion

Contributions

We recall that the purpose of this thesis was to study the geographical and socio-demographic disparities in oncology care pathways, in metropolitan France.

In the first chapter, we described the hospitals available in the country, and characterized them regarding oncology specialization. That characterization process was automatically performed through an unsupervised clustering algorithm, trained on hospitals statistics from the SAE public survey. We were then able to differentiate the most suited hospitals for oncology care, and isolate the hospitals that had no oncology activity. Then, we studied the collaborations between these hospitals, measured by the number of patients who visited a common hospital during their pathways. From this collaboration dataset, we could discover communities of hospitals that frequently exchange patients together. These communities contain hospitals with different degree of oncology specialization. This information could be a starting point to creating oncology collaboration groups, consisting in hospitals working together to make sure the hospitals with less expertise are continuously trained by more specialized hospitals.

In the next chapter, we studied the accessibility to oncology care centers in metropolitan France. We computed an accessibility score for every municipality in metropolitan France. The score reflects how easy it would be for patients from a given municipality to reach an oncology specialized hospital. This score is based on a weighting between supply and demand, as well as travel impedance. We described the spatial distribution of this score, which was higher in dense areas, near the most specialized hospitals, identified through the clustering

step.

Then, we proposed an optimization algorithm to identify which hospitals to grow in order to maximize the oncology accessibility. This algorithm took as input the current accessibility distribution, as well as some user-defined constraints. Such constraints may include a maximum hospital growth percentage, based on the current hospital oncology specialization. Through this optimization process, we identified a list of hospitals that should be grown in priority to improve the oncology accessibility distribution. The results were detailed for every region. We packaged our method into a web application, that could be used by healthcare professionals to run simulations and eventually improve the healthcare planning, benefiting millions of patients.

The previous work on oncology accessibility did not directly studied the actual cancer patients routes. In the next chapter, we extracted all the visited hospitals during the pathways of cancer patients, and described the duration and distance traveled based on the patients residence. These results validated our oncology accessibility score since travel durations were longer in areas with low accessibility scores. Longer travels were shown to have a negative impact on the patients prognosis and treatment. Moreover, long travels often increases patients fatigue, due to the travel burden. We argued that travel duration was not the only factor to consider when studying the tediousness of a journey. We built a composite indicator to reflect the travel burden of a route, based on duration, distance and road sinuosity. We showed that patients living in rural areas had higher travel burden, due to the longer drives they experienced, as well as the lower road quality and higher sinuosity. Finally, we proposed an algorithm that simulates a setup where every patient would visit the closest specialized hospital, while making sure the hospitals capacities were not exceeded. We showed that this approach could reduce the average driving duration by 36%, as well as the associated carbon footprint of the journey.

Although, in practice, patients are oriented to an hospital by their general practitioner. There are multiple evidences in the literature that patients are not satisfied with the level of information they receive during their pathways. In cancer care, some patients could be

sent to the wrong hospital, without them noticing. When that is the case, the hospital could either be a well suited hospital, but unnecessarily far from the patient residence; or an hospital that is not experienced enough in the patients pathology. For these reasons, we built "healthcare-network", a web application that lists all the hospitals in metropolitan France, and displays key statistics on them. The application could be used by patients to learn more about the hospitals around them, and by health professionals, to make sure the hospital they are sending their patients are well suited for their pathologies. We believe such tool could incentivize physicians to send patients closer to their location of residence. Moreover, bringing more transparency to oncology care could be a way to reduce disparities, provided that all the population has an equal access to these online tools.

Future work

Our oncology specialization clusters could be used in further research to assess whether the oncology care pathways are more often degraded in hospitals from the least specialized clusters. For instance, our clusters could be the input variables of survival analyses, to assess whether there are significant variations in the prognosis based on the oncology specialization of the chosen hospital. More research could also be done on the effectiveness of collaborations between the oncology communities we discovered. These communities are a first proposition of hospitals candidates that could work together to better treat patients in the neighboring municipalities. Similarly, our accessibility scores could be used in survival analyses, to assess whether patients living in the areas with low accessibility scores have more degraded pathways and lower prognosis. Regarding the web applications we developed, they could be introduced to healthcare professionals in France, like the Regional Health Agencies (ARS), responsible of the organization and the coordination of the hospitals in the country. Working closely with these professionals would allow to adapt our tools to their needs, so they can eventually be used in practice to take concrete decisions on the planning of care in the country.

Bibliography

[1] Ferlay J, Colombet M, Soerjomataram I, Parkin DM, Piñeros M, Znaor A, et al. Cancer statistics for the year 2020: An overview. International Journal of Cancer. 2021 Apr;.

[2] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA: A Cancer Journal for Clinicians. 2021;71(3):209–249. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21660. Available from: https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21660.

[3] Vanhaecht K. The impact of clinical pathways on the organisation of care processes; 2007. Book Title: The impact of clinical pathways on the organisation of care processes. Available from: https://lirias.kuleuven.be/retrieve/92842.

[4] Schrijvers G, van Hoorn A, Huiskes N. The care pathway: concepts and theories: an introduction. International Journal of Integrated Care. 2012 Sep;12(Special Edition Integrated Care Pathways):e192. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3602959/.

[5] Zon RT, Frame JN, Neuss MN, Page RD, Wollins DS, Stranne S, et al. American Society of Clinical Oncology Policy Statement on Clinical Pathways in Oncology. Journal of Oncology Practice. 2016 Mar;12(3):261–266.

[6] Adler NE, Newman K. Socioeconomic Disparities In Health: Pathways And Policies. Health Affairs. 2002 Mar;21(2):60–76. Publisher: Health Affairs. Available from: https://www.healthaffairs.org/doi/10.1377/hlthaff.21.2.60.

[7] Lubega J, Kimutai RL, Chintagumpala MM. Global health disparities in childhood cancers. Current Opinion in Pediatrics. 2021 Feb;33(1):33–39.

[8] Carethers JM, Doubeni CA. Causes of Socioeconomic Disparities in Colorectal Cancer and Intervention Framework and Strategies. Gastroenterology. 2020 Jan;158(2):354–367.

[9] Silber JH, Rosenbaum PR, Ross RN, Reiter JG, Niknam BA, Hill AS, et al. Disparities in Breast Cancer Survival by Socioeconomic Status Despite Medicare and Medicaid Insurance. The Milbank Quarterly. 2018 Dec;96(4):706–754.

[10] Grabinski VF, Brawley OW. Disparities in Breast Cancer. Obstetrics and Gynecology Clinics of North America. 2022 Mar;49(1):149–165.

[11] Ward E, Jemal A, Cokkinides V, Singh GK, Cardinez C, Ghafoor A, et al. Cancer disparities by race/ethnicity and socioeconomic status. CA: a cancer journal for clinicians. 2004 Apr;54(2):78–93.

[12] Kish JK, Yu M, Percy-Laurry A, Altekruse SF. Racial and ethnic disparities in cancer survival by neighborhood socioeconomic status in Surveillance, Epidemiology, and End Results (SEER) Registries. Journal of the National Cancer Institute Monographs. 2014 Nov;2014(49):236–243.

[13] Ahmad J, Muthyala A, Kumar A, Dani SS, Ganatra S. Disparities in Cardio-oncology: Effects On Outcomes and Opportunities for Improvement. Current Cardiology Reports. 2022 Sep;24(9):1117–1127.

[14] McGinnis JM, Foege WH. Actual causes of death in the United States. JAMA. 1993 Nov;270(18):2207–2212.

[15] Ferrari M, Flora N, Anderson KK, Haughton A, Tuck A, Archie S, et al. Gender differences in pathways to care for early psychosis. Early Intervention in Psychiatry. 2018 Jun;12(3):355–361.

[16] Bugiardini R, Ricci B, Cenko E, Vasiljevic Z, Kedev S, Davidovic G, et al. Delayed Care and Mortality Among Women and Men With Myocardial Infarction. Journal of the American Heart Association. 2017 Aug;6(8):e005968.

[17] Hogan J, Couchoud C, Bonthuis M, Groothoff Jw, Jager Kj, Schaefer F, et al. Gender Disparities in Access to Pediatric Renal Transplantation in Europe: Data From the ESPN/ERA-EDTA Registry. American journal of transplantation : official journal of the American Society of Transplantation and the American Society of Transplant Surgeons. 2016 Jul;16(7). Publisher: Am J Transplant. Available from: https://pubmed.ncbi.nlm.nih.gov/26783738/.

[18] Park A, Alabaster A, Shen H, Mell Lk, Katzel Ja. Undertreatment of women with locoregionally advanced head and neck cancer. Cancer. 2019 Jan;125(17). Publisher: Cancer. Available from: https://pubmed.ncbi.nlm.nih.gov/31090932/.

[19] Gong IY, Tan NS, Ali SH, Lebovic G, Mamdani M, Goodman SG, et al. Temporal Trends of Women Enrollment in Major Cardiovascular Randomized Clinical Trials. The Canadian Journal of Cardiology. 2019 May;35(5):653–660.

[20] Carter Paulson E, Wirtalla C, Armstrong K, Mahmoud Nn. Gender influences treatment and survival in colorectal cancer surgery. Diseases of the colon and rectum. 2009 Dec;52(12). Publisher: Dis Colon Rectum. Available from: https://pubmed.ncbi.nlm.nih.gov/19959975/.

[21] Rose TL, Deal AM, Nielsen ME, Smith AB, Milowsky MI. Sex disparities in use of chemotherapy and survival in patients with advanced bladder cancer. Cancer. 2016 Jul;122(13):2012–2020.

[22] Buzyn A. Le Plan cancer 2014-2019 : un plan de lutte contre les inégalités et les pertes de chance face à la maladie. Les Tribunes de la sante. 2014 Jul;nř 43(2):53–60. Bibliographie_available: 0 Cairndomain: www.cairn.info Cite Par_available: 1 Publisher: Presses de Sciences Po. Available from: https://www.cairn.info/revue-les-tribunes-de-la-sante1-2014-2-page-53.htm?ref=doi.

[23] Kelly C, Hulme C, Farragher T, Clarke G. Are differences in travel time or distance to healthcare for adults in global north countries associated with an impact on health outcomes? A systematic review. BMJ open. 2016 Nov;6(11):e013059.

[24] Pekala KR, Yabes JG, Bandari J, Yu M, Davies BJ, Sabik LM, et al. The centralization of bladder cancer care and its implications for patient travel distance. Urologic Oncology: Seminars and Original Investigations. 2021 Dec;39(12):834.e9–834.e20. Available from: https://www.sciencedirect.com/science/article/pii/S1078143921001873.

[25] Birkmeyer JD, Stukel TA, Siewers AE, Goodney PP, Wennberg DE, Lucas FL. Surgeon Volume and Operative Mortality in the United States. New England Journal of Medicine. 2003 Nov;349(22):2117–2127. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsa035205. Available from: https://doi.org/10.1056/NEJMsa035205.

[26] Finks JF, Osborne NH, Birkmeyer JD. Trends in Hospital Volume and Operative Mortality for High-Risk Surgery. New England Journal of Medicine. 2011 Jun;364(22):2128–2137. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMsa1010705. Available from: https://doi.org/10.1056/NEJMsa1010705.

[27] Hollenbeck BK, Daignault S, Dunn RL, Gilbert S, Weizer AZ, Miller DC. Getting under the hood of the volume-outcome relationship for radical cystectomy. The Journal of Urology. 2007 Jun;177(6):2095–2099; discussion 2099.

[28] Goossens-Laan CA, Gooiker GA, Gijn Wv, Post PN, Bosch JL, Kil PJ, et al. A systematic review and meta-analysis of the relationship between hospital/surgeon volume and outcome for radical cystectomy: an update for the ongoing debate. Centre for Reviews and Dissemination (UK); 2011. Publication Title: Database of Abstracts of Reviews of Effects (DARE): Quality-assessed Reviews [Internet]. Available from: https://www.ncbi.nlm.nih.gov/books/NBK81189/.

[29] Woo YL, Kyrgiou M, Bryant A, Everett T, Dickinson HO. Centralisation of services for gynaecological cancer. The Cochrane Database of Systematic Reviews. 2012 Mar;(3):CD007945.

[30] Olaitan A, Weeks J, Mocroft A, Smith J, Howe K, Murdoch J. The surgical management of women with ovarian cancer in the south west of England. British Journal of Cancer. 2001 Dec;85(12):1824–1830. Number: 12 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/6692196.

[31] Wang F. Measurement, Optimization, and Impact of Health Care Accessibility: A Methodological Review. Annals of the Association of American Geographers Association of American Geographers. 2012;102(5):1104–1112.

[32] Khan AA. An integrated approach to measuring potential spatial access to health care services. Socio-Economic Planning Sciences. 1992 Oct;26(4):275–287.

[33] Guagliardo MF. Spatial accessibility of primary care: concepts, methods and challenges. International Journal of Health Geographics. 2004 Feb;3(1):3. Available from: https://doi.org/10.1186/1476-072X-3-3.

[34] Zahnd WE, Josey MJ, Schootman M, Eberth JM. Spatial accessibility to colonoscopy and its role in predicting late-stage colorectal cancer. Health Services Research. 2021;56(1):73–83. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1475-6773.13562. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/1475-6773.13562.

[35] Alahmadi K, Al-Zahrani A, Al-Ahmadi S. Spatial Accessibility to Cancer Care Facilities in Saudi Arabia; 2013. .

[36] Launay L, Guillot F, Gaillard D, Medjkane M, Saint-Gérand T, Launoy G, et al. Methodology for building a geographical accessibility health index throughout metropolitan France. PLOS ONE. 2019 Aug;14(8):e0221417. Publisher: Public Library of Science. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0221417.

[37] Gusmano MK, Weisz D, Rodwin VG, Lang J, Qian M, Bocquier A, et al. Disparities in access to health care in three French regions. Health Policy (Amsterdam, Netherlands). 2014 Jan;114(1):31–40.

[38] Gao F, Kihal W, Le Meur N, Souris M, Deguen S. Assessment of the spatial accessibility to health professionals at French census block level. International Journal for Equity in Health. 2016 Aug;15(1):125. Available from: https://doi.org/10.1186/s12939-016-0411-z.

[39] Wang F. Why Public Health Needs GIS: A Methodological Overview. Annals of GIS. 2020;26(1):1–12.

[40] Church RL. Location modelling and GIS. Geographical information systems. 1999;1:293–303. Publisher: John Wiley Chichester, Sussex.

[41] Luo J. Integrating the Huff Model and Floating Catchment Area Methods to Analyze Spatial Access to Healthcare Services. Transactions in GIS. 2014;18(3):436–448. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/tgis.12096. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/tgis.12096.

[42] Tao Z, Cheng Y, Dai T, Rosenberg MW. Spatial optimization of residential care facility locations in Beijing, China: maximum equity in accessibility. International Journal of Health Geographics. 2014 Sep;13(1):33. Available from: https://doi.org/10.1186/1476-072X-13-33.

[43] Krugman P. Opinion | Why Inequality Matters. The New York Times. 2013 Dec;Available from: https://www.nytimes.com/2013/12/16/opinion/krugman-why-inequality-matters.html.

[44] Meyer D. Equity and efficiency in regional policy. Periodica Mathematica Hungarica. 2008 Mar;56(1):105–119. Available from: https://doi.org/10.1007/s10998-008-5105-x.

[45] Culyer AJ, Wagstaff A. Equity and equality in health and health care. Journal of Health Economics. 1993 Dec;12(4):431–457.

[46] Hemenway D. The Optimal Location of Doctors. New England Journal of Medicine. 1982 Feb;306(7):397–401. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJM198202183060704. Available from: https://doi.org/10.1056/NEJM198202183060704.

[47] Fried C. Rights and health care–beyond equity and efficiency. The New England journal of medicine. 1975;.

[48] Oliver A, Mossialos E. Equity of access to health care: Outlining the foundations for action. Journal of epidemiology and community health. 2004 Sep;58:655–8.

[49] Murad A, Faruque F, Naji A, Tiwari A. Using the location-allocation P-median model for optimising locations for health care centres in the city of Jeddah City, Saudi Arabia. Geospatial Health. 2021 Oct;16(2).

[50] Shavandi H, Mahlooji H. A fuzzy queuing location model with a genetic algorithm for congested systems. Applied Mathematics and Computation - AMC. 2006 Oct;181:440–456.

[51] Casado S, Laguna M, Pacheco J. Heuristical labour scheduling to optimize airport passenger flows. Journal of the Operational Research Society. 2005 Jun;56(6):649–658. Available from: https://www.tandfonline.com/doi/full/10.1057/palgrave.jors.2601859.

[52] Wang F, Tang Q. Planning toward Equal Accessibility to Services: A Quadratic Programming Approach. Environment and Planning B: Planning and Design. 2013 Apr;40(2):195–212. Publisher: SAGE Publications Ltd STM. Available from: https://doi.org/10.1068/b37096.

[53] Luo J, Tian L, Luo L, Yi H, Wang F. Two-Step Optimization for Spatial Accessibility Improvement: A Case Study of Health Care Planning in Rural China. BioMed Research International. 2017 Apr;2017:e2094654. Publisher: Hindawi. Available from: https://www.hindawi.com/journals/bmri/2017/2094654/.

[54] Li X, Wang F, Yi H. A two-step approach to planning new facilities towards equal accessibility. Environment and Planning B: Urban Analytics and City Science. 2017 Nov;44(6):994–1011. Publisher: SAGE Publications Ltd STM. Available from: https://doi.org/10.1177/0265813516657083.

[55] Luo W, Qi Y. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. Health & Place. 2009 Dec;15(4):1100–1107. Available from: http://www.sciencedirect.com/science/article/pii/S1353829209000574.

[56] Hanna TP, King WD, Thibodeau S, Jalink M, Paulin GA, Harvey-Jones E, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. BMJ. 2020 Nov;371:m4087. Publisher: British Medical Journal Publishing Group Section: Research. Available from: https://www.bmj.com/content/371/bmj.m4087.

[57] Caplan LS, Helzlsouer KJ. Delay in breast cancer: a review of the literature. Public Health Reviews. 1992;20(3-4):187–214.

[58] Williams F. Assessment of Breast Cancer Treatment Delay Impact on Prognosis and Survival: a Look at the Evidence from Systematic Analysis of the Literature. Journal of Cancer Biology & Research. 2015;3(4):1071.

[59] Pace LE, Mpunga T, Hategekimana V, Dusengimana JMV, Habineza H, Bigirimana JB, et al. Delays in Breast Cancer Presentation and Diagnosis at Two Rural Cancer Referral Centers in Rwanda. The Oncologist. 2015 Jul;20(7):780–788.

[60] Flytkjær Virgilsen L, Møller H, Vedsted P. Cancer diagnostic delays and travel distance to health services: A nationwide cohort study in Denmark. Cancer Epidemiology. 2019 Apr;59:115–122.

[61] Weiss DJ, Nelson A, Vargas-Ruiz CA, Gligori K, Bavadekar S, Gabrilovich E, et al. Global maps of travel time to healthcare facilities. Nature Medicine. 2020 Dec;26(12):1835–1838. Number: 12 Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41591-020-1059-1.

[62] Payne S, Jarrett N, Jeffs D. The impact of travel on cancer patients' experiences of treatment: a literature review. European Journal of Cancer Care. 2000 Dec;9(4):197–203.

[63] Ambroggi M, Biasini C, Del Giovane C, Fornari F, Cavanna L. Distance as a Barrier to Cancer Diagnosis and Treatment: Review of the Literature. The Oncologist. 2015 Dec;20(12):1378–1385.

[64] Dutta S, Biswas N, Muhkherjee G. Evaluation of Socio-demographic Factors for Non-compliance to Treatment in Locally Advanced Cases of Cancer Cervix in a Rural Medical College Hospital in India. Indian Journal of Palliative Care. 2013;19(3):158–165. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3853394/.

[65] Guidry JJ, Aday LA, Zhang D, Winn RJ. Transportation as a barrier to cancer treatment. Cancer Practice. 1997 Dec;5(6):361–366.

[66] Schroen AT, Brenin DR, Kelly MD, Knaus WA, Slingluff CL. Impact of patient distance to radiation therapy on mastectomy use in early-stage breast cancer patients. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2005 Oct;23(28):7074–7080.

[67] Celaya MO, Rees JR, Gibson JJ, Riddle BL, Greenberg ER. Travel distance and season of diagnosis affect treatment choices for women with early-stage breast cancer in a predominantly rural population (United States). Cancer causes & control: CCC. 2006 Aug;17(6):851–856.

[68] Voti L, Richardson LC, Reis IM, Fleming LE, Mackinnon J, Coebergh JWW. Treatment of local breast carcinoma in Florida: the role of the distance to radiation therapy facilities. Cancer. 2006 Jan;106(1):201–207.

[69] Meden T, St John-Larkin C, Hermes D, Sommerschield S. Relationship Between Travel Distance and Utilization of Breast Cancer Treatment in Rural Northern Michigan. JAMA. 2002 Jan;287(1):111. Available from: https://doi.org/10.1001/jama.287.1.111-JMS0102-5-1.

[70] Nattinger AB, Kneusel RT, Hoffmann RG, Gilligan MA. Relationship of distance from a radiotherapy facility and initial breast cancer treatment. Journal of the National Cancer Institute. 2001 Sep;93(17):1344–1346.

[71] Boscoe FP, Johnson CJ, Henry KA, Goldberg DW, Shahabi K, Elkin EB, et al. Geographic proximity to treatment for early stage breast cancer and likelihood of mastectomy. Breast (Edinburgh, Scotland). 2011 Aug;20(4):324–328.

[72] Satasivam P, O'Neill S, Sivarajah G, Sliwinski A, Kaiser C, Niall O, et al. The dilemma of distance: patients with kidney cancer from regional Australia present at a more advanced stage. BJU international. 2014 Mar;113 Suppl 2:57–63.

[73] Tracey E, McCaughan B, Badgery-Parker T, Young J, Armstrong BK. Patients with localized non-small cell lung cancer miss out on curative surgery with distance from specialist care. ANZ Journal of Surgery. 2015;85(9):658–663. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ans.12855. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/ans.12855.

[74] Virgilsen LF, Møller H, Vedsted P. Travel distance to cancer-diagnostic facilities and tumour stage. Health & Place. 2019 Nov;60:102208.

[75] Charlton M, Schlichting J, Chioreso C, Ward M, Vikas P. Challenges of Rural Cancer Care in the United States. Oncology (Williston Park, NY). 2015 Sep;29(9):633–640.

[76] Dasgupta P, Baade PD, Youlden DR, Garvey G, Aitken JF, Wallington I, et al. Variations in outcomes by residential location for women with breast cancer: a systematic review. BMJ open. 2018 Apr;8(4):e019050.

[77] Hall SE, Holman CDJ, Hendrie DV, Spilsbury K. Unequal access to breast-conserving surgery in Western Australia 1982-2000. ANZ journal of surgery. 2004 Jun;74(6):413–419.

[78] Tracey E, Hacker NF, Young J, Armstrong BK. Effects of access to and treatment in specialist facilities on survival from epithelial ovarian cancer in Australian women: A data linkage study. International Journal of Gynecologic Cancer. 2014 Sep;24(7):1232–1240. Available from: http://www.scopus.com/inward/record.url?scp=84906847096&partnerID=8YFLogxK.

[79] Thongsuksai P, Chongsuvivatwong V, Sriplung H. Delay in breast cancer care: a study in Thai women. Medical Care. 2000 Jan;38(1):108–114.

[80] Baade PD, Dasgupta P, Aitken JF, Turrell G. Distance to the closest radiotherapy facility and survival after a diagnosis of rectal cancer in Queensland. The Medical Journal of Australia. 2011 Sep;195(6):350–354.

[81] Lee B, Goktepe O, Hay K, Connors JM, Sehn LH, Savage KJ, et al. Effect of Place of Residence and Treatment on Survival Outcomes in Patients With Diffuse Large B-Cell Lymphoma in British Columbia. The Oncologist. 2014 Mar;19(3):283–290. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3958457/.

[82] Brundisini F, Giacomini M, DeJean D, Vanstone M, Winsor S, Smith A. Chronic disease patients' experiences with accessing health care in rural and remote areas: a systematic review and qualitative meta-synthesis. Ontario Health Technology Assessment Series. 2013;13(15):1–33.

[83] Change IIPoC, Intergovernmental Panel on Climate Change I. Climate Change 2014: Synthesis report. Geneva: IPCC; 2015. Available from: http://www.ipcc.ch/report/ar5/syr/.

[84] Pichler PP, Jaccard IS, Weisz U, Weisz H. International comparison of health care carbon footprints. Environmental Research Letters. 2019 May;14(6):064004. Publisher: IOP Publishing. Available from: https://doi.org/10.1088/1748-9326/ab19e1.

[85] Health Care Without Harm (HCWH), ARUP. The Global Road Map for Health Care Decarbonization; 2021. Available from: https://www.arup.com/en/perspectives/publications/research/section/healthcare-without-harm.

[86] Watts N, Amann M, Arnell N, Ayeb-Karlsson S, Beagley J, Belesova K, et al. The 2020 report of The Lancet Countdown on health and climate change: responding to converging crises. Lancet (London, England). 2021 Jan;397(10269):129–170.

[87] Andrews E, Pearson D, Kelly C, Stroud L, Rivas Perez M. Carbon footprint of patient journeys through primary care: a mixed methods approach. The British Journal of General Practice: The Journal of the Royal College of General Practitioners. 2013 Sep;63(614):e595–603.

[88] Nicolet J, Mueller Y, Paruta P, Boucher J, Senn N. What is the carbon footprint of primary care practices? A retrospective life-cycle analysis in Switzerland. Environmental Health: A Global Access Science Source. 2022 Jan;21(1):3.

[89] Forner D, Purcell C, Taylor V, Noel CW, Pan L, Rigby MH, et al. Carbon footprint reduction associated with a surgical outreach clinic. Journal of Otolaryngology - Head & Neck Surgery = Le Journal D'oto-Rhino-Laryngologie Et De Chirurgie Cervico-Faciale. 2021 Apr;50(1):26.

[90] Eskander A, Goldstein DP, Irish JC. Health Services Research and Regionalization of CareFrom Policy to Practice: the Ontario Experience in Head and Neck Cancer. Current Oncology Reports. 2016 Feb;18(3):19. Available from: https://doi.org/10.1007/s11912-016-0500-6.

[91] Guillon S, Nguyen Ba E, Oufkir N, Hequet D, Rouzier R. Empreinte carbone et canceră: lheure de la green oncologyă? Bulletin du Cancer. 2020 May;107(5):612–613. Available from: https://www.sciencedirect.com/science/article/pii/S0007455120301454.

[92] The Shift Project. Plan de Transformation de l'Economie Française (PTEF); 2021. Available from: https://theshiftproject.org/plan-de-transformation-de-leconomie-francaise-focus-sur-la-sante/.

[93] Mooi JK, Whop LJ, Valery PC, Sabesan SS. Teleoncology for indigenous patients: the responses of patients and health workers. The Australian Journal of Rural Health. 2012 Oct;20(5):265–269.

[94] Sabesan S, Kelly J. Are teleoncology models merely about avoiding long distance travel for patients? European Journal of Cancer Care. 2014 Nov;23(6):745–749.

[95] Sabesan S, Roberts LJ, Aiken P, Joshi A, Larkins S. Timely access to specialist medical oncology services closer to home for rural patients: experience from the Townsville Teleoncology Model. The Australian Journal of Rural Health. 2014 Aug;22(4):156–159.

[96] Sabesan S. Medical models of teleoncology: current status and future directions. Asia-Pacific Journal of Clinical Oncology. 2014 Sep;10(3):200–204.

[97] Sabesan S, Larkins S, Evans R, Varma S, Andrews A, Beuttner P, et al. Telemedicine for rural cancer care in North Queensland: bringing cancer care home. The Australian Journal of Rural Health. 2012 Oct;20(5):259–264.

[98] Bertucci F, Le Corroller-Soriano AG, Monneur-Miramon A, Moulin JF, Fluzin S, Maraninchi D, et al. Outpatient Cancer Care Delivery in the Context of E-Oncology: A French Perspective on "Cancer outside the Hospital Walls". Cancers. 2019 Feb;11(2):E219.

[99] Viswanath K, Nagler RH, Bigman-Galimore CA, McCauley MP, Jung M, Ramanadhan S. The communications revolution and health inequalities in the 21st century: implications for cancer control. Cancer Epidemiology, Biomarkers & Prevention: A Publication of the American Association for Cancer Research, Cosponsored by the American Society of Preventive Oncology. 2012 Oct;21(10):1701–1708.

[100] Viswanath K. Science and society: the communications revolution and cancer control. Nature Reviews Cancer. 2005 Oct;5(10):828–835.

[101] Butow PN, Maclean M, Dunn SM, Tattersall MH, Boyer MJ. The dynamics of change: cancer patients' preferences for information, involvement and support. Annals of Oncology: Official Journal of the European Society for Medical Oncology. 1997 Sep;8(9):857–863.

[102] Cassileth BR, Zupkis RV, Sutton-Smith K, March V. Information and participation preferences among cancer patients. Annals of Internal Medicine. 1980 Jun;92(6):832–836.

[103] Johnson J. The effects of a patient education course on persons with a chronic illness. Cancer Nursing. 1982 Apr;5(2):117–123.

[104] Hack TF, Pickles T, Bultz BD, Degner LF, Katz A, Davison BJ. Feasibility of an Audiotape Intervention for Patients with Cancer. Journal of Psychosocial Oncology. 1999 Sep;17(2):1–15. Publisher: Routledge _eprint: https://doi.org/10.1300/J077v17n02_01. Available from: https://doi.org/10.1300/J077v17n02_01.

[105] Mohide EA, Whelan TJ, Rath D, Gafni A, Willan AR, Czukar D, et al. A randomised trial of two information packages distributed to new cancer patients before their initial appointment at a regional cancer centre. British Journal of Cancer. 1996 Jun;73(12):1588–1593.

[106] McPherson CJ, Higginson IJ, Hearn J. Effective methods of giving information in cancer: a systematic literature review of randomized controlled trials. Journal of Public Health Medicine. 2001 Sep;23(3):227–234.

[107] SheaBudgell MA, Kostaras X, Myhill KP, Hagen NA. Information Needs and Sources of Information for Patients during Cancer Follow-Up. Current Oncology. 2014 Aug;21(4):165–173. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute. Available from: https://www.mdpi.com/1718-7729/21/4/1932.

[108] Huchcroft S, Snodgrass T, Troyan S, Wares C. Testing the Effectiveness of an Information Booklet for Cancer Patients. Journal of Psychosocial Oncology. 1984 Dec;2(2):73–83. Publisher: Routledge _eprint: https://doi.org/10.1300/J077v02n02_06. Available from: https://doi.org/10.1300/J077v02n02_06.

[109] Cegala DJ. Patient communication skills training: a review with implications for cancer patients. Patient Education and Counseling. 2003 May;50(1):91–94. Available from: https://www.sciencedirect.com/science/article/pii/S0738399103000879.

[110] Finney Rutten LJ, Agunwamba AA, Wilson P, Chawla N, Vieux S, Blanch-Hartigan D, et al. Cancer-Related Information Seeking Among Cancer Survivors: Trends Over a Decade (2003-2013). Journal of Cancer Education: The Official Journal of the American Association for Cancer Education. 2016 Jun;31(2):348–357.

[111] Hesse BW, Nelson DE, Kreps GL, Croyle RT, Arora NK, Rimer BK, et al. Trust and sources of health information: the impact of the Internet and its implications for health care providers: findings from the first Health Information National Trends Survey. Archives of Internal Medicine. 2005 Dec;165(22):2618–2624.

[112] Ley P. Communicating with patients: Improving communication, satisfaction and compliance. Communicating with patients: Improving communication, satisfaction and compliance. New York, NY, US: Croom Helm; 1988. Pages: xviii, 210.

[113] Hogbin B, Fallowfield L. Getting it taped: the 'bad news' consultation with cancer patients. British Journal of Hospital Medicine. 1989 Apr;41(4):330–333.

[114] Stewart MA. Effective physician-patient communication and health outcomes: a review. CMAJ: Canadian Medical Association journal = journal de l'Association medicale canadienne. 1995 May;152(9):1423–1433.

[115] Bartlett EE, Grayson M, Barker R, Levine DM, Golden A, Libber S. The effects of physician communications skills on patient satisfaction; recall, and adherence. Journal of Chronic Diseases. 1984;37(9-10):755–764.

[116] Higginson I, Wade A, McCarthy M. Palliative care: views of patients and their families. BMJ (Clinical research ed). 1990 Aug;301(6746):277–281.

[117] Anderson JG, Rainey MR, Eysenbach G. The impact of CyberHealthcare on the physician-patient relationship. Journal of Medical Systems. 2003 Feb;27(1):67–84.

[118] Dolce MC. The Internet as a source of health information: experiences of cancer survivors and caregivers with healthcare providers. Oncology Nursing Forum. 2011 May;38(3):353–359.

[119] Arora NK, Johnson P, Gustafson DH, McTavish F, Hawkins RP, Pingree S. Barriers to information access, perceived health competence, and psychosocial health outcomes: test of a mediation model in a breast cancer sample. Patient Education and Counseling. 2002 May;47(1):37–46.

[120] Chen X, Siu LL. Impact of the media and the internet on oncology: survey of cancer patients and oncologists in Canada. Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology. 2001 Dec;19(23):4291–4297.

[121] Pereira JL, Koski S, Hanson J, Bruera ED, Mackey JR. Internet usage among women with breast cancer: an exploratory study. Clinical Breast Cancer. 2000 Jul;1(2):148–153; discussion 154–155.

[122] Ziebland S, Chapple A, Dumelow C, Evans J, Prinjha S, Rozmovits L. How the internet affects patients' experience of cancer: a qualitative study. BMJ (Clinical research ed). 2004 Mar;328(7439):564.

[123] Carlsson ME. Cancer patients seeking information from sources outside the health care system: change over a decade. European Journal of Oncology Nursing: The Official Journal of European Oncology Nursing Society. 2009 Sep;13(4):304–305.

[124] Renzulli P, Lowy A, Maibach R, Egeli RA, Metzger U, Laffer UT. The influence of the surgeon's and the hospital's caseload on survival and local recurrence after colorectal cancer surgery. Surgery. 2006 Mar;139(3):296–304.

[125] Bonati E, Dell'Abate P, Rubini P, Del Rio P. Surgeon case volume and 5 years survival rate for colorectal cancer. Annali Italiani Di Chirurgia. 2021;92:654–659.

[126] McDermott AM, Wall DM, Waters PS, Cheung S, Sibbering M, Horgan K, et al. Surgeon and breast unit volume-outcome relationships in breast cancer surgery and treatment. Annals of Surgery. 2013 Nov;258(5):808–813; discussion 813–814.

[127] Yen TWF, Laud PW, Sparapani RA, Nattinger AB. Surgeon specialization and use of sentinel lymph node biopsy for breast cancer. JAMA surgery. 2014 Feb;149(2):185–192.

[128] Renzulli P, Laffer UT. Learning curve: the surgeon as a prognostic factor in colorectal cancer surgery. Recent Results in Cancer Research Fortschritte Der Krebsforschung Progres Dans Les Recherches Sur Le Cancer. 2005;165:86–104.

[129] Glaser LM, Brennan L, King LP, Milad MP. Surgeon Volume in Benign Gynecologic Surgery: Review of Outcomes, Impact on Training, and Ethical Contexts. Journal of Minimally Invasive Gynecology. 2019 Feb;26(2):279–287.

[130] Nietz S, Ruff P, Chen WC, O'Neil DS, Norris SA. Quality indicators for the diagnosis and surgical management of breast cancer in South Africa. Breast (Edinburgh, Scotland). 2020 Dec;54:187–196.

[131] Linden G, Smith B, York J. Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Computing. 2003 Jan;7(1):76–80. Conference Name: IEEE Internet Computing.

[132] Koren Y, Bell R, Volinsky C. Matrix Factorization Techniques for Recommender Systems. Computer. 2009 Aug;42(8):30–37. Conference Name: Computer.

[133] Christakopoulou K, Kawale J, Banerjee A. Recommendation with Capacity Constraints. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management. Singapore Singapore: ACM; 2017. p. 1439–1448. Available from: https://dl.acm.org/doi/10.1145/3132847.3133034.

[134] Zhao S, King I, Lyu MR. A Survey of Point-of-interest Recommendation in Location-based Social Networks. arXiv:160700647 [cs]. 2016 Jul;ArXiv: 1607.00647. Available from: http://arxiv.org/abs/1607.00647.

[135] Mandel JE, Morel-Ovalle L, Boas FE, Ziv E, Yarmohammadi H, Deipolyi A, et al. Optimizing Travel Time to Outpatient Interventional Radiology Procedures in a Multi-Site Hospital System Using a Google Maps Application. Journal of Digital Imaging. 2018 Oct;31(5):591–595.

[136] Jia T, Tao H, Qin K, Wang Y, Liu C, Gao Q. Selecting the optimal healthcare centers with a modified P-median model: a visual analytic perspective. International Journal of Health Geographics. 2014 Oct;13:42.

[137] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825–2830. Available from: http://jmlr.org/papers/v12/pedregosa11a.html.

[138] Xu D, Tian Y. A Comprehensive Survey of Clustering Algorithms. Annals of Data Science. 2015 Jun;2(2):165–193. Available from: https://doi.org/10.1007/s40745-015-0040-1.

[139] Luxburg UV. A Tutorial on Spectral Clustering; 2007.

[140] Fortunato S. Community detection in graphs. Physics Reports. 2010 Feb;486(3):75–174. Available from: https://www.sciencedirect.com/science/article/pii/S0370157309002841.

[141] Hamilton WL, Ying R, Leskovec J. Representation Learning on Graphs: Methods and Applications. arXiv; 2018. ArXiv:1709.05584 [cs]. Available from: http://arxiv.org/abs/1709.05584.

[142] Perozzi B, Al-Rfou R, Skiena S. DeepWalk: Online Learning of Social Representations. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining; 2014. p. 701–710. ArXiv:1403.6652 [cs]. Available from: http://arxiv.org/abs/1403.6652.

[143] Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. arXiv:160902907 [cs, stat]. 2017 Feb;ArXiv: 1609.02907. Available from: http://arxiv.org/abs/1609.02907.

[144] Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. arXiv; 2013. ArXiv:1310.4546 [cs, stat]. Available from: http://arxiv.org/abs/1310.4546.

[145] Kipf TN, Welling M. Variational Graph Auto-Encoders. arXiv:161107308 [cs, stat]. 2016 Nov;ArXiv: 1611.07308. Available from: http://arxiv.org/abs/1611.07308.

[146] Kingma DP, Welling M. Auto-Encoding Variational Bayes. arXiv; 2014. ArXiv:1312.6114 [cs, stat]. Available from: http://arxiv.org/abs/1312.6114.

[147] van der Maaten L, Hinton G. Viualizing data using t-SNE. Journal of Machine Learning Research. 2008 Nov;9:2579–2605.

[148] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. KDD'96. Portland, Oregon: AAAI Press; 1996. p. 226–231.

[149] Schonfeld HK, Heston JF, Falk IS. Numbers of physicians required for primary medical care. The New England Journal of Medicine. 1972 Mar;286(11):571–576.

[150] on Graduate Medical Education C. Physician Distribution and Health Care Challenges in Rural and Inner-city Areas: Council on Graduate Medical Education Tenth Report. U.S. Department of Health and Human Services, Public Health Service, Health Resources and Services Administration; 1998. Google-Books-ID: vcw5xluPwo4C.

[151] Connor RA, Hillson SD, Krawelski JE. Competition, professional synergism, and the geographic distribution of rural physicians. Medical Care. 1995 Nov;33(11):1067–1078.

[152] Connor RA, Kralewski JE, Hillson SD. Measuring geographic access to health care in rural areas. Medical Care Review. 1994;51(3):337–377.

[153] Basu J. Border-crossing adjustment and personal health care spending by state. Health Care Financing Review. 1996;18(1):215–236.

[154] Basu J, Lazenby HC, Levit KR. Medicare spending by state: the border-crossing adjustment. Health Care Financing Review. 1995;17(2):219–241.

[155] Holahan J, Zuckerman S. Border crossing for physician services: implications for controlling expenditures. Health Care Financing Review. 1993;15(1):101–122.

[156] Openshaw S. The modifiable areal unit problem. Norwick [Norfolk: Geo Books; 1983. OCLC: 12052482.

[157] Fryer GE, Drisko J, Krugman RD, Vojir CP, Prochazka A, Miyoshi TJ, et al. Multi-method assessment of access to primary medical care in rural Colorado. The Journal of Rural Health: Official Journal of the American Rural Health Association and the National Rural Health Care Association. 1999;15(1):113–121.

[158] Reilly WJ. The Law of Retail Gravitation. W.J. Reilly; 1931. Google-Books-ID: 5o9CAAAAIAAJ.

[159] Hansen WG. How Accessibility Shapes Land Use. Journal of the American Institute of Planners. 1959 May;25(2):73–76. Publisher: Routledge _eprint: https://doi.org/10.1080/01944365908978307. Available from: https://doi.org/10.1080/01944365908978307.

[160] Joseph AE, Bantock PR. Measuring potential physical accessibility to general practitioners in rural areas: A method and case study. Social Science & Medicine. 1982 Jan;16(1):85–90. Available from: https://www.sciencedirect.com/science/article/pii/0277953682904282.

[161] Talen E, Anselin L. Assessing spatial equity: An evaluation of measures of accessibility to public playgrounds. Environment and Planning A. 1998 Apr;30(4):595–613. Available from: http://www.scopus.com/inward/record.url?scp=0031816545&partnerID=8YFLogxK.

[162] Luo W. Using a GIS-based floating catchment method to assess areas with shortage of physicians. Health & Place. 2004 Mar;10(1):1–11. Available from: http://www.sciencedirect.com/science/article/pii/S1353829202000679.

[163] Langford M, Higgs G, Fry R. Multi-modal two-step floating catchment area analysis of primary health care accessibility. Health & Place. 2016 Mar;38:70–81. Available from: http://www.sciencedirect.com/science/article/pii/S1353829216000022.

[164] Wan N, Zou B, Sternberg T. A three-step floating catchment area method for analyzing spatial access to health services. International Journal of Geographical Information Science. 2012 Jun;26(6):1073–1089. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/13658816.2011.624987. Available from: https://doi.org/10.1080/13658816.2011.624987.

[165] Huff DL. A Probabilistic Analysis of Shopping Center Trade Areas. Land Economics. 1963;39(1):81–90. Publisher: [Board of Regents of the University of Wisconsin System, University of Wisconsin Press]. Available from: https://www.jstor.org/stable/3144521.

[166] Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020 Mar;17(3):261–272. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 3 Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Biophysical chemistry;Computational biology and bioinformatics;Technology Subject_term_id: biophysical-chemistry;computational-biology-and-bioinformatics;technology. Available from: https://www.nature.com/articles/s41592-019-0686-2.

[167] Bertsimas D, Tsitsiklis J. Introduction to Linear Optimization; 1998. Journal Abbreviation: Athena Scientific Publication Title: Athena Scientific.

[168] Monge G. Mémoire sur la théorie des déblais et des remblais. Imprimerie royale; 1781. Google-Books-ID: IG7CGwAACAAJ.

[169] Kantorovitch L. On the Translocation of Masses. Management Science. 1958 Oct;5(1):1–4. Publisher: INFORMS. Available from: https://pubsonline.informs.org/doi/abs/10.1287/mnsc.5.1.1.

[170] Cuturi M. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. arXiv:13060895 [stat]. 2013 Jun;ArXiv: 1306.0895. Available from: http://arxiv.org/abs/1306.0895.

[171] Knopp P, Sinkhorn R. Concerning nonnegative matrices and doubly stochastic matrices. Pacific Journal of Mathematics. 1967 Jan;21(2):343–348. Publisher: Pacific Journal of Mathematics, A Non-profit Corporation. Available from: https://projecteuclid.org/journals/pacific-journal-of-mathematics/volume-21/issue-2/Concerning-nonnegative-matrices-and-doubly-stochastic-matrices/pjm/1102992505.full.

[172] Peyré G, Cuturi M. Computational Optimal Transport. arXiv:180300567 [stat]. 2020 Mar;ArXiv: 1803.00567. Available from: http://arxiv.org/abs/1803.00567.

[173] Flamary R, Courty N, Gramfort A, Alaya MZ, Boisbunon A, Chambon S, et al. POT: Python Optimal Transport. Journal of Machine Learning Research. 2021;22(78):1–8. Available from: http://jmlr.org/papers/v22/20-451.html.

[174] Narducci F, Musto C, Polignano M, de Gemmis M, Lops P, Semeraro G. A Recommender System for Connecting Patients to the Right Doctors in the HealthNet Social Network. In: Proceedings of the 24th International Conference on World Wide Web. WWW '15

Companion. New York, NY, USA: Association for Computing Machinery; 2015. p. 81–82. Available from: https://doi.org/10.1145/2740908.2742748.

[175] Hoens TR, Blanton M, Chawla NV. Reliable medical recommendation systems with patient privacy. In: Proceedings of the 1st ACM International Health Informatics Symposium. IHI '10. New York, NY, USA: Association for Computing Machinery; 2010. p. 173–182. Available from: https://doi.org/10.1145/1882992.1883018.

[176] Han Q, Ji M, Martinez de Rituerto de Troya I, Gaur M, Zejnilovic L. A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care. In: 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA); 2018. p. 481–490.

[177] Tran TNT, Felfernig A, Trattner C, Holzinger A. Recommender systems in the healthcare domain: state-of-the-art and research issues. Journal of Intelligent Information Systems. 2021 Aug;57(1):171–201. Available from: https://doi.org/10.1007/s10844-020-00633-6.

[178] Zhang Y, Chen M, Huang D, Wu D, Li Y. iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. Future Generation Computer Systems. 2017 Jan;66:30–35. Available from: https://www.sciencedirect.com/science/article/pii/S0167739X15003842.

[179] Calero Valdez A, Ziefle M, Verbert K, Felfernig A, Holzinger A. Recommender Systems for Health Informatics: State-of-the-Art and Future Perspectives. Lecture Notes in Computer Science. 2016 Nov;9605.

[180] Viswanath K, Ackerson LK. Race, ethnicity, language, social class, and health communication inequalities: a nationally-representative cross-sectional study. PloS One. 2011 Jan;6(1):e14550.

[181] Li X, Cong G, Li XL, Pham TAN, Krishnaswamy S. Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In: Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '15. New York, NY, USA: Association for Computing Machinery; 2015. p. 433–442. Available from: https://doi.org/10.1145/2766462.2767722.

[182] Genevay A, Peyré G, Cuturi M. Learning Generative Models with Sinkhorn Divergences. arXiv:170600292 [stat]. 2017 Oct;ArXiv: 1706.00292. Available from: http://arxiv.org/abs/1706.00292.

[183] Cuturi M, Blondel M. Soft-DTW: a Differentiable Loss Function for Time-Series. arXiv:170301541 [stat]. 2018 Feb;ArXiv: 1703.01541. Available from: http://arxiv.org/abs/1703.01541.

[184] Tai KS, Bailis P, Valiant G. Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training. arXiv:210208622 [cs, stat]. 2021 Jun;ArXiv: 2102.08622. Available from: http://arxiv.org/abs/2102.08622.

[185] Dupuy A, Galichon A, Sun Y. Estimating matching affinity matrix under low-rank constraints. arXiv:161209585 [stat]. 2016 Dec;ArXiv: 1612.09585. Available from: http://arxiv.org/abs/1612.09585.

List of Figures

List of Tables

université PARIS-SACLAY

ÉCOLE DOCTORALE
Cancérologie : biologie - médecine - santé (CBMS)

Titre: Etude des disparités géographiques et socio-démographiques dans les parcours de soins en oncologie.

Mots clés: 3 à 6 mots clefs (version en français)

Résumé: On estime à 382 000 le nombre de nouveaux cas de cancers incidents et à 157 400 le nombre de décès en 2018 en France. Le Plan Cancer 2014-2019 annonce les objectifs à mettre en oeuvre dans la lutte contre le cancer en France. En particulier, les objectifs 2 et 7 insistent sur la qualité du parcours de soin : ils ambitionnent respectivement de "garantir la qualité et la sécurité des prises en charge" et "assurer des prises en charge globales et personnalisés". Dans le but de standardiser le parcours de soin tout en personnalisant la prise en charge, des trajectoires de soin ont été instaurées. La définition de ces trajectoires de soin optimales s'appuie sur des recommandations de bonnes pratiques nationales et internationales. Nous proposons d'étudier en détails les disparités géographiques et socio-démographiques dans les parcours de soin des patients atteints d'un cancer en France. Dans un premier temps, nous chercherons à caractériser les établissements de santé en France à partir de leur activité en oncologie. Puis, nous étudierons la distribution de ces centres sur le territoire, afin de mettre en évidence d'éventuelles disparités dans l'accès à ceux-ci. Enfin, nous tenterons de proposer un algorithme de recommandations de centres de soins, à partir des données du Système National des Données de Santé (SNDS). Cet algorithme aura pour but de guider les patients vers le centre optimal, en vue de maximiser la qualité du parcours de soins.

Title: Geographic and socio-demographic disparities in oncology care pathways.

Keywords: 3 à 6 mots clefs (version en anglais)

Abstract: In France, during year 2018, there was 382,000 new cancer cases and 157,400 deaths. The 2014-2019 Cancer Plan sets new objectives for cancer care in France. In particular, objectives 2 and 7 emphasize the quality of the care pathways: they aim respectively to guarantee the quality and safety of care and ensure comprehensive and personalized care. In order to standardize the care pathways while personalizing care, care trajectories have been introduced. The definition of these optimal care trajectories is based on national and international good practice recommendations. We propose to study in detail the geographical and socio-demographic disparities in the care pathways of cancer patients in France. First, we will try to characterize the health care institutions in France based on their oncology activity. Then, we will study the distribution of these centers on the territory, in order to highlight possible disparities in access to them. Finally, we will try to propose a care center recommendation algorithm, using the French social security database (SNDS). This algorithm will aim at guiding patients towards the optimal center, in order to maximize the quality of the care pathways.

CAMION: a catchment area maximization algorithm, with application to oncology accessibility in metropolitan France.

Eric Daoud[1,2], Anne-Sophie Hamy[1,3], Elise Dumas[1,4,5], Lidia Delrieu[1], Beatriz Grandal Rejo[1,6], Christine Le Bihan-Benjamin[7], Sophie Houzard[7], Philippe-Jean Bousquet[8,9], Judicaël Hotton[10], Aude-Marie Savoye[11], Christelle Jouannaud[11], Chloé-Agathe Azencott[4,5,12], Marc Lelarge[2], Fabien Reyal[1,6,10,*]

1. Residual Tumor & Response to Treatment Laboratory, RT2Lab, INSERM, U932 Immunity and Cancer, Institut Curie, Université Paris, 75005 Paris, France
2. INRIA, DI/ENS, PSL Research University, Paris, France
3. Department of Medical Oncology, Institut Curie, Paris, France
4. INSERM, U900, 75005 Paris, France
5. MINES ParisTech, PSL Research University, CBIO-Centre for Computational Biology, 75006 Paris, France
6. Department of Surgery, Institut Curie, Paris, France
7. Health Data and Assessment Department, Survey Data Science and Assessment Division, National Cancer Institute, 52 avenue André Morizet 92100 Boulogne-Billancourt, France
8. Aix Marseille Univ, Inserm, IRD, SESSTIM, Equipe Labellisée Ligue Contre le Cancer, Marseille, France
9. Survey Data Science and Assessment Division, National Cancer Institute, 52 avenue André Morizet 92100 Boulogne-Billancourt, France
10. Department of Surgery, Institut Jean Godinot, Reims, France
11. Department of Medical Oncology, Institut Jean Godinot, Reims, France
12. Institut Curie, PSL Research University, 75005 Paris, France

*Corresponding Author: Fabien Reyal, Institut Curie, 26 rue d'Ulm, Paris 75005, France. Phone: +33(0)142346339; E-mail: fabien.reyal@curie.fr.

Abstract

Background: Access to health services plays a key role in cancer survival. Uneven distributions of populations and health facilities lead to geographical disparities. Location-allocation algorithms can address these disparities by finding new locations and capacities for health facilities. However, in oncology, opening new hospitals or moving them is difficult in practice, and should be handled carefully.

Methods: We propose a method to measure the spatial accessibility to oncology care and identify the hospitals to grow to reduce disparities. We first ran a clustering algorithm to automatically label the hospitals in terms of oncology specialization. Then, we computed an accessibility score to these hospitals for every population location. Finally, we introduced *CAMION*, an optimization algorithm based on Linear Programming that reduces disparities in oncology accessibility by identifying health facilities that should increase their capacities.

Results: We demonstrate our algorithm in metropolitan France. The clustering step let us identify different oncology specialization levels for hospitals. Most of the population in metropolitan France lived in good accessibility areas, especially in large cities. Lower accessibility zones are often rural or suburban municipalities. The optimization algorithm effectively manages to identify hospitals to grow, based on current oncology specialization and accessibility scores.

Discussion: There is a tradeoff to be found by patients, between care center proximity and care center expertise, which is less likely to happen for patients living in good accessibility areas. The accessibility score is deliberately non-specific to cancer type but can be adapted to more precise pathologies. Our method is replicable in any country, given hospitals and population locations data. We developed a web application intended for healthcare professionals to let them to run the optimization algorithm with the desired parameters and visualize the results.

Introduction

Cancer is a leading cause of death worldwide, accounting for nearly 10 million deaths in 2020. While a lot of the ongoing research is focusing on finding new cancer treatments, accessibility to oncology care receives less attention. Yet, several studies have showed that access to health services plays a key role in cancer survival. For instance, geographic residency

status and social environment seem to explain treatment and prognosis disparities for patients with non-small cell lung cancer (1). In France, increases in travel times to health services were associated with lower survival rates for patients with a colorectal cancer (2). In New Zealand, living in deprived areas, far from a cancer center or from primary care was associated with lower survival chances for patients with colorectal, lung and prostate cancers (3).

Accessibility refers to the relative ease by which services can be reached from a given location (4). Accessibility can be defined by spatial factors, determined by where you are; and non-spatial factors, determined by who you are (5). Spatial accessibility methods assess the availability of supply locations from demand locations, connected by a travel impedance metric. Supply locations are characterized by their capacity or quantity of available resource. Similarly, demand locations are characterized by their population. Such methods have been successfully used to measure access to healthcare, such as primary care (6) or oncology care (4,7,8) in several countries including France (9–11). In what follows, we restrict accessibility to spatial accessibility and use both terms interchangeably.

Uneven distributions of population and health-care providers lead to geographic disparity in accessibility for patients (12). For instance, Weiss et al. (13) showed that 8.9% of the global population could not reach healthcare within one hour if they have access to motorized transport. In Germany, Bauer et al. (14) shown that 10% of the population lived in areas with low accessibility for internal medicine and surgery. Location-allocation algorithms (15) can optimize the distribution and supply of health providers to reduce accessibility disparities. These algorithms seek the optimal placement of facilities for a desirable objective under certain constraints (4). For instance, Luo et al. developed an optimization algorithm to improve the healthcare planning in rural China by finding the best place and capacity for new health facilities (16). Tao et al. worked on a spatial optimization model to maximize equity in accessibility to residential care facility in Beijing, China (17). When optimizing health accessibil-

ity, there are two competing goals: equity and efficiency (18,19). Equity may be defined as equal access to healthcare for everyone (20). An efficient situation is when everything has been done to help any person without harming anyone else (21). While some argue that efficiency should be addressed in priority (21), others agree that equity is a matter of ethical obligation, especially in public health (22,23).

The goal of this paper is to apply spatial accessibility methods to oncology care centers and propose an optimization algorithm to reduce disparities. We demonstrate our results in metropolitan France. There are many care centers in France, which do not share the same degree of oncology specialization. Therefore, we first run a clustering algorithm to automatically group the care centers based on their medical statistics and attributes. Using these clusters, we label the care centers in terms of hospital development and oncology specialization. Then, we compute an oncology accessibility score for every municipality in metropolitan France. We then introduce *CAMION*, an optimization algorithm based on *Linear Programming* which uses the clusters of care centers and the accessibility scores to suggest, given a limited budget, where to increase hospital capacity to improve the oncology accessibility. Finally, our method is packaged into a web application intended to healthcare professionals so they can run the optimization algorithm with the desired parameters for any region.

Methods

Data collection

Health data is collected from two sources: the *French national administrative database* (PMSI) and the *French annual health facilities statistics* (*SAE*). *PMSI* data includes discharge summaries for all inpatients admitted to public and private hospitals in France. The *SAE* database is a compulsory and exhaustive administrative survey of all public and private hospitals in France. The survey is sent every year and describes the activities of the hospitals as well as the list of services and their staff. We restricted the analysis to the year 2018. We in-

clude every hospital in metropolitan France that declared a *Medicine, Surgery or Obstetric (MCO)* activity in the *SAE* survey, in 2018. We also included the liberal radiotherapy care centers, with no *MCO* activity. The resulting dataset contains 1,662 care centers.

Geographic and travel data were retrieved from open data platforms. Municipalities and their census statistics were extracted from the *National Statistics Bureau of France (INSEE)* website. We used the *OpenRouteService (ORS) API* to compute the driving routes between hospitals and municipalities, which is necessary for the accessibility score.

Care centers characterization

We selected a list of 24 variables with the help of medical experts to characterize the care centers. The list of variables and their definitions is available in the supplementary materials. The variables are either binary when they encode the presence or absence of a service; or discrete when they encode the number of stays. We only focus on treatments received in hospitals.

Given the large number of care centers, we use a clustering algorithm to automatically group together similar care centers. More specifically, we first run a *Principal Component Analysis (PCA)* algorithm on the *SAE* dataset that describes the care centers. The input data has 24 variables, and we perform the dimensionality reduction with $n = 2$ components. We tried different number of components, from 2 to 5, but we found 2 gave good and easy to interpret results. We then run a clustering algorithm on the PCA-reduced dataset to automatically isolate care centers with similar statistics. We tried several algorithms like *K-Means (24), DBSCAN (25) and Spectral Clustering* (26). In our case, *Spectral Clustering* with 8 clusters gave the most interpretable and better isolated groups. For the number k of clusters, we tested all values from 2 to 10 and manually interpreted the results with medical experts.

Accessibility score

There are several ways to compute accessibility to healthcare (6). The easiest and most straightforward methods are computed within bordered areas, like provider-to-population ratios in each municipality. While they are very intuitive, these methods do not account for border crossing, or travel impedance, which makes them less accurate. Recently, a new type of method has been developed and is now used in most spatial accessibility papers. This algorithm is called *Two Step Floating Catchment Area (2SFCA)* (27). It is a two-step method that first computes a provider-to-population ratio for each provider location. In the second step, for each population location, an accessibility score is obtained by summing the provider-to-population ratios. For the algorithm to work, a catchment threshold (distance or travel time) must be set. Above this threshold, a provider location is considered unreachable from the population location, and vice versa. The *2SFCA* method does not account for distance decay: a care center is either reachable or not. The *Enhanced Two Step Floating Catchment Area (e2SFCA)* (28) addresses this limitation by applying weights to differentiate travel zones in both steps.

We now explain more formally how to compute *eS2FCA* scores. Consider P_i the population at location i, with $1 \leq i \leq n$ where n is the number of population locations. Similarly, consider S_u the capacity of care center u, with $1 \leq u \leq m$ where m is the number of care centers. Finally, let d_{iu} be the matrix of size $n \times m$ containing the distances between location i and care center u. We consider r sub-catchment zones each associated with a weight W_s, and a distance D_s, with $1 \leq s \leq r$, such that $D_1 < D_2 < \cdots < D_r$ and $W_1 > W_2 > \cdots > W_r$. The resulting r travel intervals are $I_1 = [0, D_1]$, $I_2 = [D_1, D_2]$, … , $I_r = [D_{r-1}-, D_r]$. The accessibility A_i of a population location i is computed in two steps. Step 1: for every care center u, compute its weighted capacity-to-population ratio R_u. Step 2: for every population location, compute A_i as the sum all the weighted R_u of the reachable care centers.

$$R_u = \frac{S_u}{\sum_{s=1}^{r} W_s \sum_{i, d_{iu} \in I_s} P_i}$$

$$A_i = \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} R_u$$

The capacity of a care center is balanced by the total population with access to it. A population location that solely has access to low capacities or overcrowded care centers will have a low accessibility score. Similarly, a population location will have low accessibility scores if the distance to get to the nearby care centers is large.

As we want to compute the accessibility to oncology care centers, we chose S_u to be the oncology activity of a hospital u. We define oncology activity as the sum of the number of medical and surgery stays related to cancer, and the number of patients with chemotherapy or radiotherapy. A care center with no oncology activity will have $R_u = 0$ and a municipality that solely has access to this care center u will have $A_i = 0$. We use driving duration as travel impedance metric, and we set the maximum catchment area to a 90-minute drive. In 2018, only 24,152 patients out of 761,057 (3.2%) had travel duration greater than 90 minutes for cancer related pathways. This is low enough to consider that care centers are non-reachable beyond this distance. We divide the catchment area into 3 intervals: $I_1 = (0, 30]$, $I_2 = (30, 60]$ and $I_3 = (60, 90]$. The associated weights are respectively $W_1 = 1, W_2 = 0.042$ and $W_3 = 0.09$. These sub catchment areas are set based on the cancer pathways travel duration distributions and validated with medical experts. The weights are the same than the *e2SFCA* paper (28).

For privacy reasons, municipalities with small populations are grouped in entities called "geographic codes" in the *PMSI* data. We decided to compute the accessibility score for each geographic code and municipalities that are grouped in the same code will have the same accessibility score.

Accessibility optimization

Regarding efficiency optimization, the most popular algorithms are *p-median*, location set covering problem *(LSCP)* and *maximum covering location problem (MCLP)*. The *p-median* algorithm minimizes the weighted sum of distances between users and facilities (29). *LSCP* minimizes the number of facilities needed to cover all demand (30). *MCLP* maximizes the demand covered within a desired distance or time threshold by locating a given number of facilities (31).

To reach equal access to healthcare, quadratic programming has been used to minimize the variance of accessibility scores defined by the *2SFCA* (32). Similarly, a *Particle Swarm Optimization (PSO)* algorithm was developed to minimize the total square difference between the accessibility score of each demand location and the weighted average accessibility score (17). Finally, a two-step optimization algorithm has been developed to address the dual objectives of efficiency and equality, by first choosing where to site new hospitals and then deciding which capacity they should have (16,33).

However, most of the previous algorithms seek locations to open new health facilities. In this work, we are interested in the case where the health facilities are fixed, and the only lever to improve accessibility is to increase their capacities. Given a capacity budget, we want to know which facilities to grow and by how much. We introduce *CAMION*, an accessibility optimization algorithm based on *Floating Catchment Area* and *Linear Programming*. The initial accessibility score was computed with the *Enhanced Two Step Floating Catchment Area (e2SFCA)* (28) but our algorithm can generalize to more *FCA* derivatives.

We model the problem as an optimization task. In our case, we want our optimization algorithm to find new care centers capacities given some constraints, so that the total accessibility is maximum. We apply optimization on a given region only, rather than on the whole metropolitan France. We chose this approach because healthcare planning is handled region-

ally rather than nationally. We show below that our optimization problem is a Linear Programming problem.

In its standard form, *Linear Programming* finds a vector x that maximizes $c^T x$ under constraints $Ax \leq \mathrm{b}$, where A is a matrix and b a vector. Boundaries can be set to x such as $x \geq 0$. Consider x_u the new capacity of a care center u, to be computed by the algorithm. Let Q_u and W_u be two vectors of size m, defined as follows:

$$Q_u = \sum_{s=1}^{r} W_s \sum_{i, d_{iu} \in I_s} P_i$$

$$W_u = \sum_{s=1}^{r} \sum_{i, d_{iu} \in I_s} W_s$$

We can compute the total accessibility as a sum on the m care centers:

$$\sum_i A_i = \sum_i \sum_{s=1}^{r} W_s \sum_{u, d_{iu} \in I_s} \frac{S_u}{Q_u}$$

$$\sum_i A_i = \sum_s \sum_{i,u, d_{iu} \in I_s} W_s \frac{S_u}{Q_u}$$

$$\sum_i A_i = \sum_u \frac{S_u}{Q_u} \sum_s \sum_{i, d_{iu}} W_s$$

$$\sum_i A_i = \sum_u \frac{S_u}{Q_u} W_u$$

The last equation can be rewritten in the *Linear Programming* standard form with:

$$c = \frac{W_u}{Q_u}$$

$$x_u = S_u$$

$$b \geq \sum_u x_u$$

$$x_{u_{min}} \leq x_u \leq x_{u_{max}}$$

The user-defined parameters are b, $x_{u_{min}}$ and $x_{u_{max}}$. b is the total capacity to be shared across all the care centers. $x_{u_{min}}$ and $x_{u_{max}}$ are the capacity boundaries for care center u. If b is set to the current total capacity, a care center can't be grown unless another one is decreased. If $b > \sum_u x_u$, the capacity of care centers can be increased without decreasing other centers. We know how to solve *Linear Programming* and we used the *SciPy* (34) implementation of the revised simplex method as explained in (35).

We now detail how we set the user-defined parameters to apply the *Linear Programming* algorithm to our specific case. The additional capacity was set as +3% of the overall activity of the region's care centers: $b = 1.03 \times \sum_u x_u$. The choice of the boundaries $x_{u_{min}}$ and $x_{u_{max}}$ is crucial and must be realistic. We studied the hospitals activity on the past four years (2016 to 2019) to retrieve the average growth percentage of a care center. The growth percentage is computed as follows: $(S_{2019} - S_{2016})/S_{2016}$. Among the care centers that grew and who had an existing oncology activity, the mean growth percentage was 23%. Hence, we set $x_{u_{max}}$ as +20% of the care center capacity. Regarding $x_{u_{min}}$, we set the boundary based on the cluster of the care center. For the three most specialized clusters, we set their $x_{u_{min}}$ equal to their current activity. We did this to prevent the algorithm from decreasing the most specialized and well-equipped care centers. Regarding the care centers from the other clusters, $x_{u_{min}} = 0$, so that they could be emptied if need be. Finally, we set $x_{u_{max}} = 0$ if the care center belongs to the least specialized cluster. The new capacities are indicative and should be further investigated to make sure they are relevant. Especially when setting an existing oncology activity to 0.

We developed a web application that allows the users to run the optimization algorithm in any region with the parameters they want. The application displays accessibility results and optimization outcomes on an interactive map with additional plots. The user can browse the list of care centers by cluster and the list of municipalities with their accessibility scores.

Results

Population and hospitals distribution in metropolitan France

In 2018, the population in France was 66,993 million. Mainland France hosts 64,812 million inhabitants (96.8%), while the remaining 2,181 million (3.2%) live in overseas departments and regions. Metropolitan France is divided into 13 administrative regions and 96 departments. The population density in France is unevenly distributed. In 2020, the overall population density in metropolitan France was 119 inhabitants per square kilometer. *Ile-de-France* region has the highest population density with 1,022 inhabitants per square kilometer. Density in other regions in metropolitan France range between 40 and 187 inhabitants/km^2. Denser areas are located near the coastline and around the largest cities like *Paris, Marseille, Lyon, Strasbourg, Toulouse, or Bordeaux*. The middle of the country is rural, and the population densities are low. While there are a great variety of regions and landscapes, the country is becoming more urbanized. This "*rural exodus*" is largely responsible of what is known as the "*empty diagonal*", a band of very low-density population that stretches from the southwest to the northeast.

We now describe the spatial distribution and specificities of the 1,662 hospitals included in this study. There are different types of hospitals in France: *Centres Hospitaliers* (CH, n=667) and *Centres Hospitaliers Régionaux / Universitaires* (CHR/U, n=142) are state-run hospitals; *Centres de Lutte Contre le Cancer* (CLCC, n=26) and *Participants au Service Public Hospitalier / Etablissement à But Non Lucratif* (PSPH/EBNL, n=142) are both private hospitals of collective interest, though *CLCC* are oncology dedicated; private hospitals (n=606) are privately run and for-profit. The non-*MCO* care centers with radiotherapy activity (n=79) are mostly private practice structures and are referred as *Other*. **Table 1** shows the number of care centers and their oncology activity per hospital type and region. Most of the care centers are public, but a non-neglectable part are private. *CLCC* represent only 1.6% of

the care centers, yet they are responsible for 14.2% of the overall oncology activity. The care centers are unevenly distributed across the country. For instance, *Corse and Centre-Val-de-Loire* are the only two regions with no *CLCC* care centers. Moreover, the proportion of oncology activity per hospital type varies from a region to another. For instance, in *Nouvelle-Aquitaine*, 47.1% of the oncology activity is handled by private care centers, whereas in *Provence-Alpes-Cote-d'Azur* it is 21.4%.

Variable value per region N = number of centers A = oncology activity (radio., chemo., surgery)	**Hospital type** **CH** **n=667**	**CHR/U** **n=142**	**CLCC** **n=26**	**Other** **n=79**	**PSPH/EBNL** **n=142**	**Private** **n=606**	**Overall** **n=1,662**
Auvergne-Rhône-Alpes	98 (49,2%)	21 (10,6%)	2 (1%)	7 (3,5%)	13 (6,5%)	58 (29,1%)	199
	34,597 (26,7%)	31,706 (24,5%)	16,966 (13,1%)	6,710 (5,2%)	6,146 (4,7%)	33,297 (25,7%)	129,422
Bourgogne-Franche-Comté	53 (64,6%)	4 (4,9%)	1 (1,2%)	5 (6,1%)	2 (2,4%)	17 (20,7%)	82
	12,238 (27,6%)	10,621 (24%)	5,844 (13,2%)	4,405 (9,9%)	657 (1,5%)	10,571 (23,8%)	44,336
Bretagne	38 (33%)	8 (7%)	1 (0,9%)	6 (5,2%)	11 (9,6%)	51 (44,3%)	115
	15,953 (27%)	11,020 (18,6%)	6,341 (10,7%)	5,553 (9,4%)	2,050 (3,5%)	18,199 (30,8%)	59,116
Centre-Val de Loire	29 (46,8%)	4 (6,5%)	0 (0%)	6 (9,7%)	2 (3,2%)	21 (33,9%)	62
	6,989 (19,6%)	11,524 (32,2%)	0 (0%)	5,137 (14,4%)	32 (0,1%)	12,058 (33,7%)	35,74
Corse	7 (53,8%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	6 (46,2%)	13
	3,486 (66,3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1,773 (33,7%)	5,259
Grand Est	70 (41,7%)	17 (10,1%)	3 (1,8%)	6 (3,6%)	30 (17,9%)	42 (25%)	168
	17,428 (19,6%)	22,123 (24,9%)	13,176 (14,8%)	6,793 (7,7%)	7,683 (8,7%)	21,553 (24,3%)	88,756
Hauts-de-France	56 (40%)	11 (7,9%)	1 (0,7%)	11 (7,9%)	12 (8,6%)	49 (35%)	140
	21,864 (26%)	15,934 (19%)	6,947 (8,3%)	8,618 (10,3%)	5,242 (6,2%)	25,399 (30,2%)	84,004
Île-de-France	40 (47,6%)	5 (6%)	4 (4,8%)	6 (7,1%)	3 (3,6%)	26 (31%)	84
	7,573 (14,9%)	7,947 (15,7%)	14,210 (28%)	5,419 (10,7%)	0 (0%)	15,627 (30,8%)	50,776
Normandie	70 (44,9%)	10 (6,4%)	1 (0,6%)	7 (4,5%)	12 (7,7%)	56 (35,9%)	156
	37,844 (33%)	26,244 (22,9%)	7,477 (6,5%)	7,157 (6,2%)	2,824 (2,5%)	33,271 (29%)	114,817
Nouvelle-Aquitaine	66 (37,7%)	14 (8%)	2 (1,1%)	7 (4%)	6 (3,4%)	80 (45,7%)	175
	14,735 (12,1%)	20,915 (17,2%)	16,047 (13,2%)	11,572 (9,5%)	1,098 (0,9%)	57,374 (47,1%)	121,741
Occitanie	34 (44,7%)	5 (6,6%)	3 (3,9%)	4 (5,3%)	5 (6,6%)	25 (32,9%)	76
	11,901 (18,9%)	11,374 (18,1%)	12,564 (19,9%)	3,422 (5,4%)	3,916 (6,2%)	19,822 (31,5%)	62,999
Pays de la Loire	53 (34,4%)	10 (6,5%)	3 (1,9%)	4 (2,6%)	15 (9,7%)	69 (44,8%)	154
	14,632 (13,6%)	16,533 (15,4%)	21,924 (20,4%)	6,172 (5,7%)	10,918 (10,2%)	37,176 (34,6%)	107,355
Provence-Alpes-Côte d'Azur	53 (22,3%)	33 (13,9%)	5 (2,1%)	10 (4,2%)	31 (13%)	106 (44,5%)	238
	24,390 (12,6%)	66,406 (34,2%)	34,028 (17,5%)	12,817 (6,6%)	14,981 (7,7%)	41,577 (21,4%)	194,199
Grand Total	667 (40,1%)	142 (8,5%)	26 (1,6%)	79 (4,8%)	142 (8,5%)	606 (36,5%)	1662
	223,630 (20,4%)	252,347 (23%)	155,524 (14,2%)	83,775 (7,6%)	55,547 (5,1%)	32,7697 (29,8%)	1,098,520

Table 1: Number of care centers (N) and overall oncology activity (A) per hospital type and region. Oncology activity is the sum of the number of patients with radiotherapy or chemotherapy, and the number of medical or surgery stays related to cancer. *CH* and *CHR/U* are public hospitals; *CLCC* and *PSPH/EBNL* are private hospitals of collective interest, though *CLCC* are oncology dedicated; private hospitals are for-profit. *Other* hospitals are mostly private practice *radiotherapy* structures. The percentages sum to 100% row-wise. In *Nouvelle-Aquitaine* region, 47.1% of the oncology activity is handled by private care centers, whereas in *Provence-Alpes-Cote-d'Azur* region it is 21.4%.

Care centers characterization

While it is obvious that *CLCC* care centers are suited for oncology care, it is difficult to assess the degree of oncology specialization for other care centers. Our clustering algorithm assigns the n=1,662 care centers into 8 clusters, sorted by oncology specialization. **Figure 1** shows the distribution of some of the key health services per cluster. These services are *biology, radiotherapy, chemotherapy, cancer surgery, intensive unit, palliative care, oncology unit, medication circuit, surgery, and outpatient surgery*. The three oncology services are *cancer surgery, radiotherapy, and chemotherapy*. We see that care centers from clusters 1 (n=79) and 2 (n=39) all have these 3 services, hence they are the most suited hospitals for oncology care. Centers from cluster 3 (n=451) have *cancer surgery* and *chemotherapy* but lack *radiotherapy*. The most part of the n=381 centers from cluster 4 have *cancer surgery*, but no *radiotherapy* nor *chemotherapy*. Care centers from cluster 5 (n=2) and cluster 6 (n=7) have *radiotherapy* and *chemotherapy* services, but no *cancer surgery*. Care centers in cluster 7 (n=77) are dedicated to *radiotherapy* and mostly private practice structures. Finally, care centers 8 (n=626) have none of the 3 oncology services. To sum up, hospitals from clusters 1 and 2 (n=118) are "all-in-one" care centers that provide the most "ideal" oncology care. Centers from clusters 3 and 4 (n=382) provide oncology care but will have to be coordinated with

additional structures during the pathways. Hospitals within clusters 5, 6 and 7 (n=86) are not allowed to perform *cancer surgery* but provide *chemotherapy* or *radiotherapy*. The remaining n=626 care centers in cluster 8 are not equipped for oncology care. Hospital types are unevenly distributed among the clusters. For instance, 76.9% of the *CLCC* care centers are placed in cluster 1, as they are the most specialized centers. In cluster 7, we find external *radiotherapy* units of some CLCC centers, and private practice structures. The proportion of private care centers varies as well: cluster 1 has almost no private care center while cluster 2 has 61.5% of private hospitals. Moreover, most of the oncology activity is handled by care centers from clusters 1 and 3. Also, the overall oncology activity from the n=79 centers in cluster 1 is almost as large as the activity of the n=451 hospitals from cluster 4.

Figure 1: Distribution of the care centers services and equipment per cluster. Each radar plot axis shows the percentage of the care centers within the cluster that have the corresponding attribute. In Cluster 1, the care centers have all the listed services. In cluster 8, the centers have almost none of the services. Care centers from cluster 1 (n=79) and cluster 2 (n=39) are the most suited for oncology care.

Accessibility score computation

We computed the spatial accessibility score to these care centers for every municipality in metropolitan France, using the *e2SFCA* algorithm and oncology activity as supply variable. We compared the accessibility distributions with *e2SFCA* vs. regular *2SFCA*. The accessibility was lower with *e2SFCA* because of the weight decay. We also studied the influence of the supply variable in the accessibility score. Accessibility is much higher if we use the number of *Medical, Surgery and Obstetric* (*MCO*) stays as supply, instead of the oncology activity. This makes sense since oncology care centers are less common and the overall *MCO* activity is higher than the oncology activity. The oncology accessibility is unevenly distributed across

the country, as displayed on **Figure 2**. For better readability, we cut the accessibility scores into 5 quantiles. *Q5* colored in dark green contains the top 20% accessibility municipalities, and *Q1* in light yellow contains the bottom 20% ones. The lowest accessibility zones are mostly located in the center of the country and in mountainous regions like the *Alps* or the *Pyrenees*. Plot (B) shows that most of the population (51.6%) lives in top 20% accessibility municipalities, while 6.3 % lives in the bottom 20% quantile. On map (A), care centers are displayed as squares, colored by cluster index, and sized by oncology activity. We see that accessibility is highest near the most specialized care centers. Indeed, the proportion of care centers from specialized clusters decreases in lower accessibility quantiles (C). We then ranked the departments by median accessibility and showed the top-10 and bottom-10 on plot (D). Among the top-5 departments, 4 are in *Ile-de-France*. Departments from the bottom-10 are rural or mountainous areas like *Lozère* and *Alpes-de-Haute-Provence*. We notice disparities within departments as well, as outlined by the large interquartile range in *Hérault* or *Alpes-Maritimes*. On the contrary, this spread is very narrow in *Ile-de-France* departments.

Figure 2: Distribution of the accessibility score computed with enhanced two step floating catchment area (e2SFCA), in metropolitan France. Plot (A) shows municipalities colored by accessibility quantile. The care centers are drawn as squares, colored by cluster, and sized by oncology activity. Plot (B) shows the total population by accessibility quantile. Plot (C) displays the percentage of care centers by cluster by accessibility quantile. Plot (D) shows the top 10 and bottom 10 list of the departments, ranked by median accessibility.

Accessibility score should be put into perspective with population density. Overall, the denser municipalities have a median accessibility around 0.02. Municipalities with low population densities have more extreme values. **Figure 3** compares accessibility and population density for three different regions: *Provence-Alpes-Cote-d'Azur* (A), *Ile-de-France* (B), and *Bourgogne-Franche-Comté* (C). Municipalities are displayed as squares, colored by accessi-

bility quantile, and sized by population density. These regions show very different profiles. In *Provence-Alpes-Cote-d'Azur* (A), accessibility is essentially low in non-dense municipalities near the *Alps*. However, in *Bourgogne-Franche-Comté* (C), we see dense municipalities with poor accessibility scores, representing a large proportion of the region. We also drew similar maps (D, E and F) where municipalities are colored based on the average travel duration for patients with cancer in 2018. We see that the average travel time is higher in municipalities with poor accessibility scores. The surface percentage with low accessibility varies from a region to another. For instance, in *Bourgogne-Franche-Comté*, 34.5% of the region has a Q1 accessibility, that is 15.6% of the region's population. Sometimes, the Q1 surface can be large but might contain very few inhabitants. This happens in *Ile-de-France*, where 15% of the surface is Q1 accessibility, representing less than 1% of the region's population. Finally, we compared our accessibility score with the department exit ratio, by municipality. Department exit ratio is defined as the proportion of cancer patients who visited a care center outside from their department of residence and was computed using the *PMSI* database. In *Provence-Alpes-Cote-d'Azur*, the exit ratio is higher in departments with low accessibility scores and few oncology specialized care centers, as in *Alpes-de-Haute-Provence* and *Hautes-Alpes*. While the *Var* department has some oncology centers, exit ratio remains high since larger care centers are in *Marseille* and *Nice*.

Figure 3: Comparison of population density with accessibility scores and patient average travel time for cancer pathways. Showing results in three regions: *Provence-Alpes-Cote-d'Azur* (A, D), *Ile-de-France* (B, E) and *Bourgogne-Franche-Comté* (C, F). Municipalities are drawn as squares, sized by population density and colored by either accessibility quantile (A, B, C) or patient average travel time (D, E, F).

We now focus on the region *Provence-Alpes-Cote-d'Azur*. This region is the far south-eastern on the mainland. The region's population was 5,048 million in 2018. Its prefecture

and largest city is *Marseille*. The region contains six departments. *Bouches-du-Rhone, Var* and *Alpes-Maritimes* are located on the coastline and gather the largest cities like *Marseille*, *Nice*, or *Toulon*. *Alpes-de-Haute-Provence, Vaucluse, and Hautes-Alpes* are inland departments, with a majority of rural and mountainous areas. Results are shown on **Figure 4**. By comparing maps (A) and (B), we confirm that the accessibility is maximum in denser areas of the region. Average patients travel time are displayed on map (C) and we drew the major roads (primary, motorway, and truck) in red. The road system is well developed on the coast, rallying the larger cities of the region. However, driving from the rural areas in the *Alps* to the major cities is hard, resulting in higher travel times. The accessibility is unevenly spread within the departments, especially in *Alpes-Maritimes* where the distribution is multi-modal (D). There, cities like *Nice* and *Cannes* have large hospitals thus good accessibility, while the northern areas of the department are mostly mountains. Accessibility is higher in municipalities with dense populations, for all the departments (E). Finally, the average travel time decreases when the accessibility score increases. This makes sense since the accessibility score was computed based on the driving distance between population locations and care centers. However, it confirms that patients living in poor accessibility zones effectively travel further to seek oncology care. In *Bouches-du-Rhone,* nearly all the municipalities have an average travel time lower than 30 minutes, while in *Alpes-de-Haute-Provence,* average travel times are rarely lower than 60 minutes (F).

Figure 4: Accessibility distribution in Provence-Alpes-Cote-d'Azur region. Map (A) shows the region accessibility distribution per municipality. Map (B) displays the population density discretized in 5 bins. The map on plot (C) displays the average travel time for cancer pathways. Large roads (primary, motorway, and trucks) are drawn in red. Plot (D) shows the accessibility distribution per department of the region. Plot (E) shows the accessibility distribution by municipality population density and department. Plot (F) compares the accessibility score from municipalities with the average travel time for cancer pathways.

Accessibility optimization

Since we focused on describing the accessibility situation in *Provence-Alpes-Cote-d'Azur*, we now present the outcomes of our optimization algorithm in this same region. The algorithm was run with the user-specified parameters stated in the Methods Section: we chose to increase the overall oncology activity in the region by 3% (+3,221 activity) and capped care centers to a 20% maximum growth. The median accessibility in the region went from 0.0093 to 0.0103, a 11.1% increase. The results are shown on **Figure 5**. Map (A) displays the accessibility delta ($A_{i_{after}} - A_{i_{before}}$) as well as the care centers eligible to grow. Centers from cluster 8 were hidden since we considered that they couldn't provide any oncology activity. The algorithm identified a list of 26 care centers where the oncology activity could grow to maximize the total accessibility in the region. These centers are either public or private hospitals, primarily located in the *Avignon* and *Gap* areas. The care centers located in high accessibility areas near *Marseille* and *Nice* were ignored by the algorithm because improving these zones is not a priority. The care center that grew the most is *Clinique Sainte Catherine*, in *Avignon*. Interestingly, this care center was recently bought by the *Unicancer* group, which coordinates all the cancer centers in France. This hospital's type will change to become a new *CLCC*. Thus, it is expected to grow in the next years and to be equipped with more oncology services and staff.

Figure 5: Accessibility delta in Provence-Alpes-Cote-d'Azur (PACA) region after running the optimization algorithm. Map (A) displays the accessibility delta ($A_{i_{after}} - A_{i_{before}}$) by municipality. Plot (B) shows the capacity delta ($S_{u_{after}} - S_{u_{before}}$) distribution. Capacity was defined as the oncology activity: the number of patients with chemotherapy or radiotherapy and the number of medical or surgery stays related to oncology. We show the list of the care centers that grew the most (C) and by how much. For instance, the hospital *Institut Sainte Catherine* in *Avignon*, was assigned a +1,030 capacity, for a total of n=6,179. Additional activity was 3,221. 26 centers grew and 1 decreased. Median accessibility before optimization was 0.0093 and 0.0103 after, corresponding to a 11.1% increase. Accessibility increased around cities like *Avignon* and *Gap*. Care centers near *Nice* were left unchanged by the algorithm.

While we described the results in *Provence-Alpes-Cote-d'Azur* region, we ran the algorithm with similar parameters on every region in metropolitan France. The results are available in the *Supplementary Materials* and on the web application. We observe two types of optimization strategies. For most regions, the algorithm manages to find a couple of areas where the accessibility can be locally improved, like it did in *Provence-Alpes-Cote-d'Azur* near *Gap* and *Avignon*. However, for regions like *Ile-de-France* and *Haut-de-France*, the hospital capacity increase is more uniformly distributed across the region. Most of the time, the algorithm left untouched the large care centers located in dense cities with good accessibilities. This can be explained by the relatively low value of the additional activity parameter: with a very large value of additional activity, every care center will grow. If we keep it low, the algorithm identifies in which areas hospital capacity should be increased in priority.

Discussion

We observe disparities in both care centers and their accessibility. The clustering algorithm successfully groups similar hospitals and lets us identify the care centers best suited for oncology care. Some variables in the *SAE* survey are declarative and potentially differ from the reality. We are aware of this bias, but we do not expect major differences that could distort our clustering results.

Receiving treatment in a care center with *surgery*, *chemotherapy* and *radiotherapy* activities is easier for the patient and leads to better care pathways. Care centers from cluster 1 will be the better choice for cancer treatment and correspond to modern oncology care specifications. However, these centers are a minority and sparsely located, essentially in dense areas and in large cities. While the inhabitants of large cities and metropolitan areas will have no problem reaching them, rural areas residents live far away from these centers. This population often has better access to care centers from intermediate clusters. Such centers do not have all the key services and the patients are more likely to visit multiple hospitals during their care pathways.

Longer drives to reach a more specialized care center could be considered more acceptable for *surgery*, where the hospital volumetry and surgeon expertise matter. However, for more frequent interventions like *chemotherapy* and *radiotherapy* especially, patients should prioritize short travels. There is a tradeoff to be found by patients, between care center proximity and care center expertise. This dilemma will be more frequent for patients living in rural areas than patients living in dense cities with large care centers nearby.

Specific attention should be given to municipalities with very poor access to oncology care centers. While we saw that most of the population lives in high accessibility areas, around 6% of the population lives in the bottom 20% accessibility quantile. Among these municipalities, some are very rural and mountainous like those in the *Alpes-de-Haute-*

Provence in *Provence-Alpes-Cote-d'Azur* region. Such areas cannot be expected to have a very good healthcare coverage. By contrast, the case of suburban areas with relatively dense population and poor accessibility should be addressed more easily. Our optimization algorithm can help driving public health policies, as it effectively identifies areas where accessibility could grow, by allocating additional oncology activity to a restricted number of care centers. The proposed growth factors are indicative and do not have to be effective within a year, as it represents a considerable effort for care centers to increase their activity.

Our oncology accessibility score is deliberately non-specific to cancer type. This score is meant to outline how easy it would be for a population location to reach a first entry point for oncology care. Here, we are only focusing on surgery, chemotherapy, and radiotherapy treatments. The same technique could be used on a specific cancer type, the method will remain the same, only the supply variable used in the accessibility score will change. We should mention that spatial accessibility is better suited for pathologies that are relatively well handled across the whole country. Accessibility for rare diseases like pediatric cancer or complex cancers that require a specific expertise is less informative because only a handful of care centers are indicated.

Similarly, we could compute an accessibility score that is focused on specific kinds of stays: our web application lets the user pick between surgery, chemotherapy, or radiotherapy as supply variable.

The quality of oncology care is linked with the care centers' volumetry. A care center with a very low activity is less likely to provide decent care. As a result, the *French National Institute of Cancer (INCa)* defined several thresholds (36) that forbid care centers with very low activity to keep operating. Similarly, the care quality in a saturated care center won't be good either, since patients are more likely to wait longer before diagnosis or between interventions. While it is easy to spot care centers with low activity, it is harder to judge if a care center is

over-crowded, and we should be careful when attributing new activity to the hospitals. We based the 20% max growth out of the previous centers' activity increase. This percentage could be tailored to the center cluster or current activity. Volumetry is not the only factor determining care quality. More sophisticated indicators like average delay between diagnosis and first treatment can tell whether a care center is in line with the care pathways recommendations. Care centers with activities lower than the thresholds, or with a large proportion of degraded pathways should be handled with care by our algorithm.

Accessibility optimization depends on many factors and healthcare professionals will not have the same uses for our algorithm. Some may consider that for a care center to grow another should decline, where others would rather not decrease any centers' activities. Moreover, the healthcare planning is very different from a region to another, and even within the regions departments are showing disparities. Hence, we cannot expect the algorithm to be used with the same parameters on every region. For all these reasons, we believe that providing a web application to run the algorithm and choose the parameters is the most useful way to the help healthcare professionals improve the current situation.

Our work is in line with the *French Cancer Plan* (37) that emphasizes the importance of increasing accessibility to oncology care as well as minimizing disparities across the country. The government mandated *INCa* to work on the accessibility development. This study and the web application we developed could help when attributing the care centers authorizations. Working closely with researchers from *INCa* and public health professionals could have a major impact on the oncology care spatial organization in metropolitan France, benefiting millions of patients.

We ran this method in metropolitan France, but it could work on any country if data on hospitals and municipalities are available.

Acknowledgements

The authors thank Thomas Ansart and the *Sciences Po* cartography workshop for helping us to improve our figures. We also thank Julien Guerin and Johan Archinard, from the *Data Factory at Institut Curie*, who helped us to deploy our web application. Finally, we thank Hakim Idjis, Marc-Felix Degni and Olivier Auliard and their team at *Capgemini Invent*, who helped us to extract the driving routes with OpenRouteService.

References

1. Johnson AM, Hines RB, Johnson JA, Bayakly AR. Treatment and survival disparities in lung cancer: the effect of social environment and place of residence. Lung Cancer Amst Neth. 2014 Mar;83(3):401–7.

2. Dejardin O, Jones AP, Rachet B, Morris E, Bouvier V, Jooste V, et al. The influence of geographical access to health care and material deprivation on colorectal cancer survival: Evidence from France and England. Health Place. 2014 Nov 1;30:36–44.

3. Haynes R, Pearce J, Barnett R. Cancer survival in New Zealand: ethnic, social and geographical inequalities. Soc Sci Med 1982. 2008 Sep;67(6):928–37.

4. Wang F. Measurement, Optimization, and Impact of Health Care Accessibility: A Methodological Review. Ann Assoc Am Geogr Assoc Am Geogr. 2012;102(5):1104–12.

5. Khan AA. An integrated approach to measuring potential spatial access to health care services. Socioecon Plann Sci. 1992 Oct;26(4):275–87.

6. Guagliardo MF. Spatial accessibility of primary care: concepts, methods and challenges. Int J Health Geogr. 2004 Feb 26;3(1):3.

7. Zahnd WE, Josey MJ, Schootman M, Eberth JM. Spatial accessibility to colonoscopy and its role in predicting late-stage colorectal cancer. Health Serv Res. 2021;56(1):73–83.

8. Alahmadi K, Al-Zahrani A, Al-Ahmadi S. Spatial Accessibility to Cancer Care Facilities in Saudi Arabia. In 2013.

9. Launay L, Guillot F, Gaillard D, Medjkane M, Saint-Gérand T, Launoy G, et al. Methodology for building a geographical accessibility health index throughout metropolitan France. PLOS ONE. 2019 août;14(8):e0221417.

10. Gusmano MK, Weisz D, Rodwin VG, Lang J, Qian M, Bocquier A, et al. Disparities in access to health care in three French regions. Health Policy Amst Neth. 2014 Jan;114(1):31–40.

11. Gao F, Kihal W, Le Meur N, Souris M, Deguen S. Assessment of the spatial accessibility to health professionals at French census block level. Int J Equity Health. 2016 Aug 2;15(1):125.

12. Wang F. Why Public Health Needs GIS: A Methodological Overview. Ann GIS. 2020;26(1):1–12.

13. Weiss DJ, Nelson A, Vargas-Ruiz CA, Gligorić K, Bavadekar S, Gabrilovich E, et al. Global maps of travel time to healthcare facilities. Nat Med. 2020 Dec;26(12):1835–8.

14. Bauer J, Klingelhöfer D, Maier W, Schwettmann L, Groneberg DA. Spatial accessibility of general inpatient care in Germany: an analysis of surgery, internal medicine and neurology. Sci Rep. 2020 Nov 5;10(1):19157.

15. Church RL. Location modelling and GIS. Geogr Inf Syst. 1999;1:293–303.

16. Luo J, Tian L, Luo L, Yi H, Wang F. Two-Step Optimization for Spatial Accessibility Improvement: A Case Study of Health Care Planning in Rural China. BioMed Res Int. 2017 Apr 18;2017:e2094654.

17. Tao Z, Cheng Y, Dai T, Rosenberg MW. Spatial optimization of residential care facility locations in Beijing, China: maximum equity in accessibility. Int J Health Geogr. 2014 Sep 1;13(1):33.

18. Krugman P. Opinion | Why Inequality Matters. The New York Times [Internet]. 2013 Dec 16 [cited 2022 May 2]; Available from: https://www.nytimes.com/2013/12/16/opinion/krugman-why-inequality-matters.html

19. Meyer D. Equity and efficiency in regional policy. Period Math Hung. 2008 Mar 1;56(1):105–19.

20. Culyer AJ, Wagstaff A. Equity and equality in health and health care. J Health Econ. 1993 Dec;12(4):431–57.

21. Hemenway D. The Optimal Location of Doctors. N Engl J Med. 1982 Feb 18;306(7):397–401.

22. Fried C. Rights and health care--beyond equity and efficiency. N Engl J Med. 1975;

23. Oliver A, Mossialos E. Equity of access to health care: Outlining the foundations for action. J Epidemiol Community Health. 2004 Sep 1;58:655–8.

24. Lloyd S. Least squares quantization in PCM. IEEE Trans Inf Theory. 1982 Mar;28(2):129–37.

25. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. Portland, Oregon: AAAI Press; 1996. p. 226–31. (KDD'96).

26. Luxburg UV. A Tutorial on Spectral Clustering. 2007.

27. Luo W. Using a GIS-based floating catchment method to assess areas with shortage of physicians. Health Place. 2004 Mar 1;10(1):1–11.

28. Luo W, Qi Y. An enhanced two-step floating catchment area (E2SFCA) method for measuring spatial accessibility to primary care physicians. Health Place. 2009 Dec 1;15(4):1100–7.

29. Murad A, Faruque F, Naji A, Tiwari A. Using the location-allocation P-median model for optimising locations for health care centres in the city of Jeddah City, Saudi Arabia. Geospatial Health. 2021 Oct 19;16(2).

30. Shavandi H, Mahlooji H. A fuzzy queuing location model with a genetic algorithm for congested systems. Appl Math Comput - AMC. 2006 Oct 1;181:440–56.

31. Casado S, Laguna M, Pacheco J. Heuristical labour scheduling to optimize airport passenger flows. J Oper Res Soc. 2005 Jun;56(6):649–58.

32. Wang F, Tang Q. Planning toward Equal Accessibility to Services: A Quadratic Programming Approach. Environ Plan B Plan Des. 2013 Apr 1;40(2):195–212.

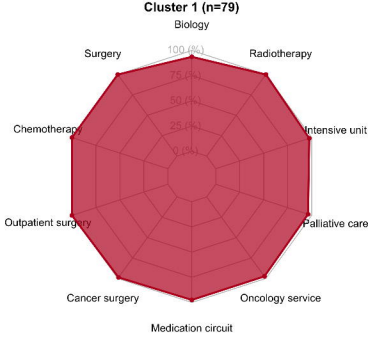
33. Li X, Wang F, Yi H. A two-step approach to planning new facilities towards equal accessibility. Environ Plan B Urban Anal City Sci. 2017 Nov 1;44(6):994–1011.

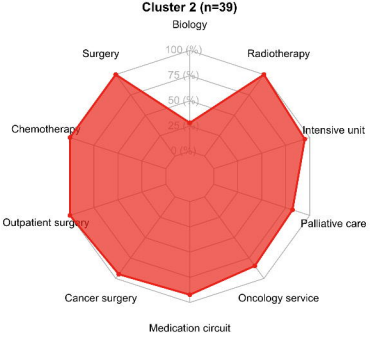
34. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods. 2020 Mar;17(3):261–72.

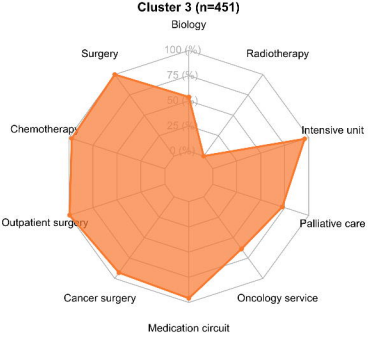
35. Bertsimas D, Tsitsiklis J. Introduction to Linear Optimization. Athena Scientific. 1998.

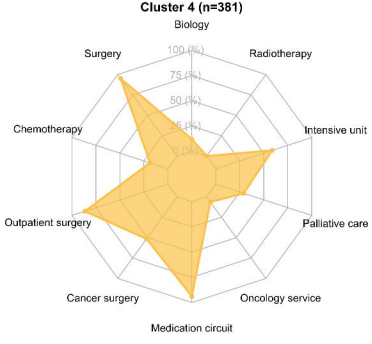
36. Institut National du Cancer. Mesure des activités soumises à seuil. 2017.

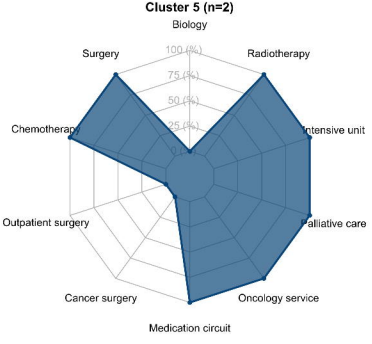
37. Buzyn A. Le Plan cancer 2014-2019 : un plan de lutte contre les inégalités et les pertes de chance face à la maladie. Trib Sante. 2014 Jul 15;n° 43(2):53–60.

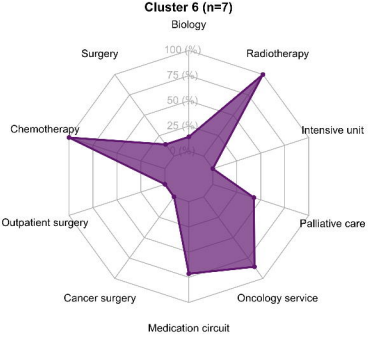

Cluster 1 (n=79)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

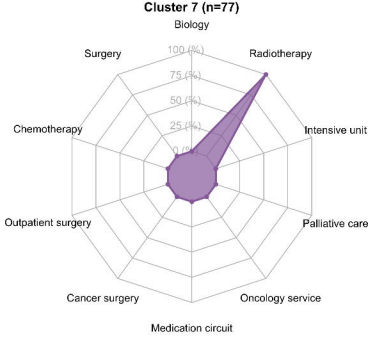

Cluster 2 (n=39)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

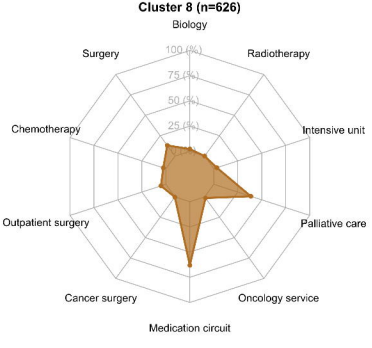

Cluster 3 (n=451)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

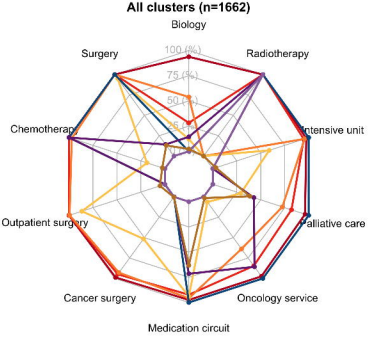

Cluster 4 (n=381)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

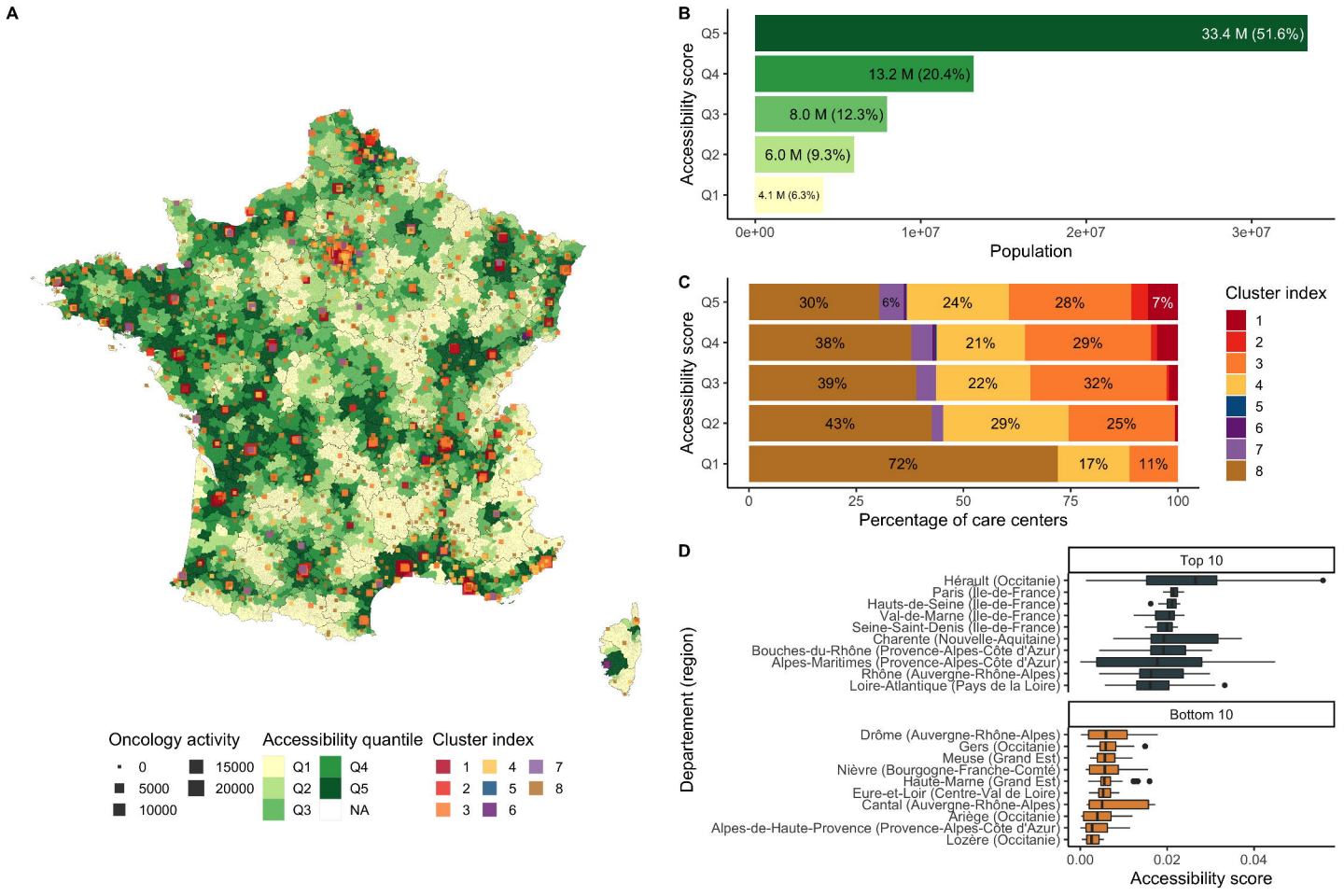

Cluster 5 (n=2)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

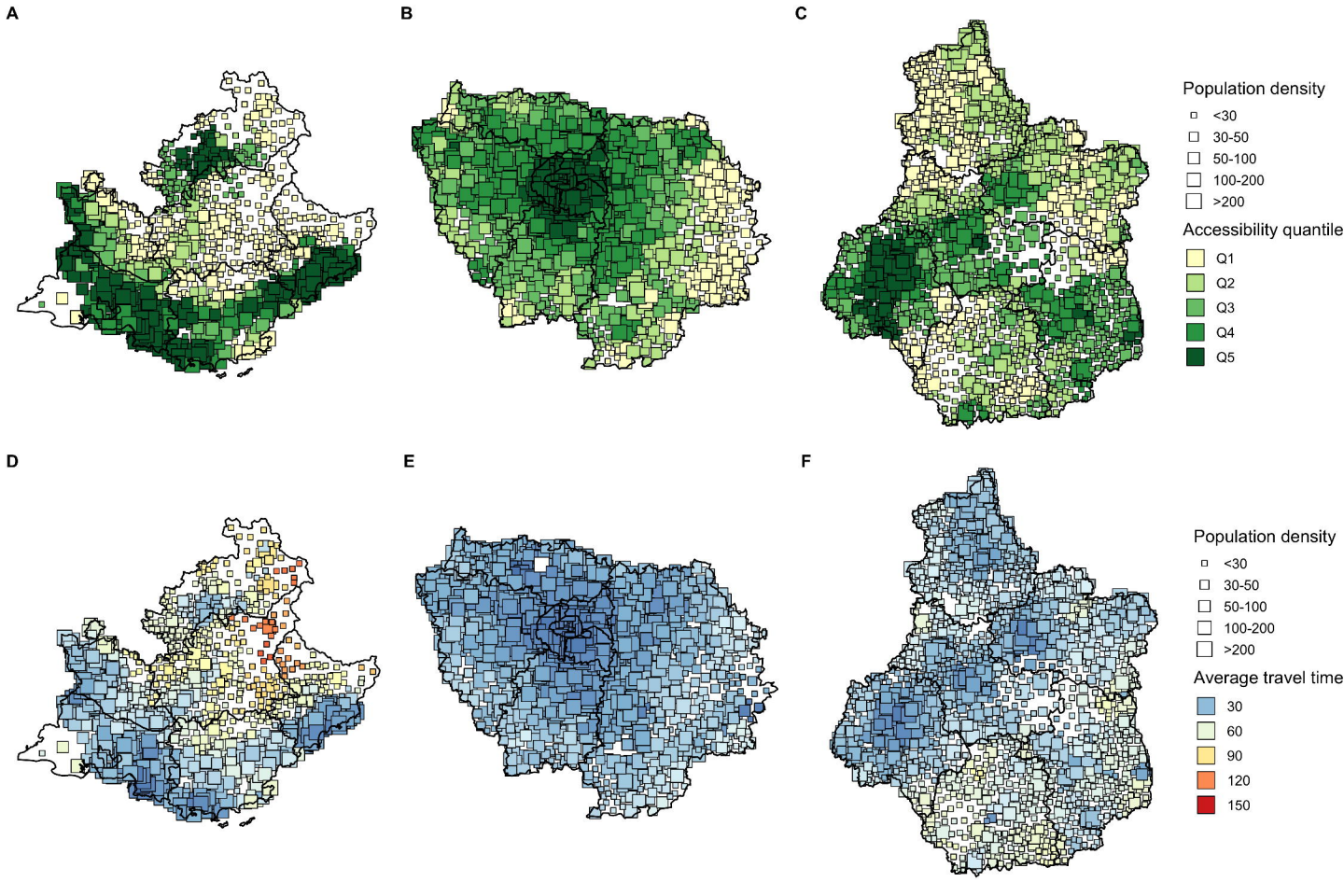

Cluster 6 (n=7)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

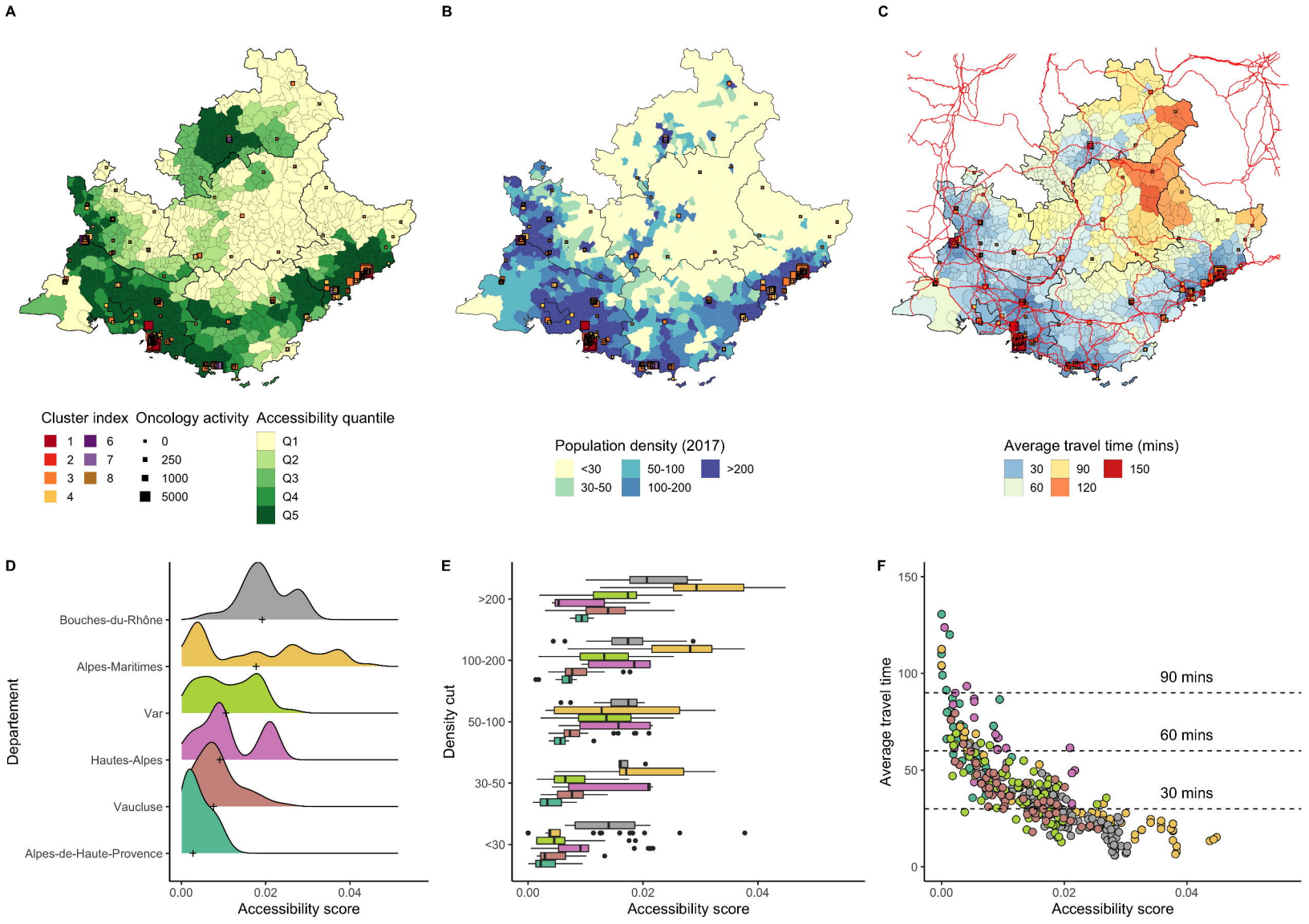

Cluster 7 (n=77)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)

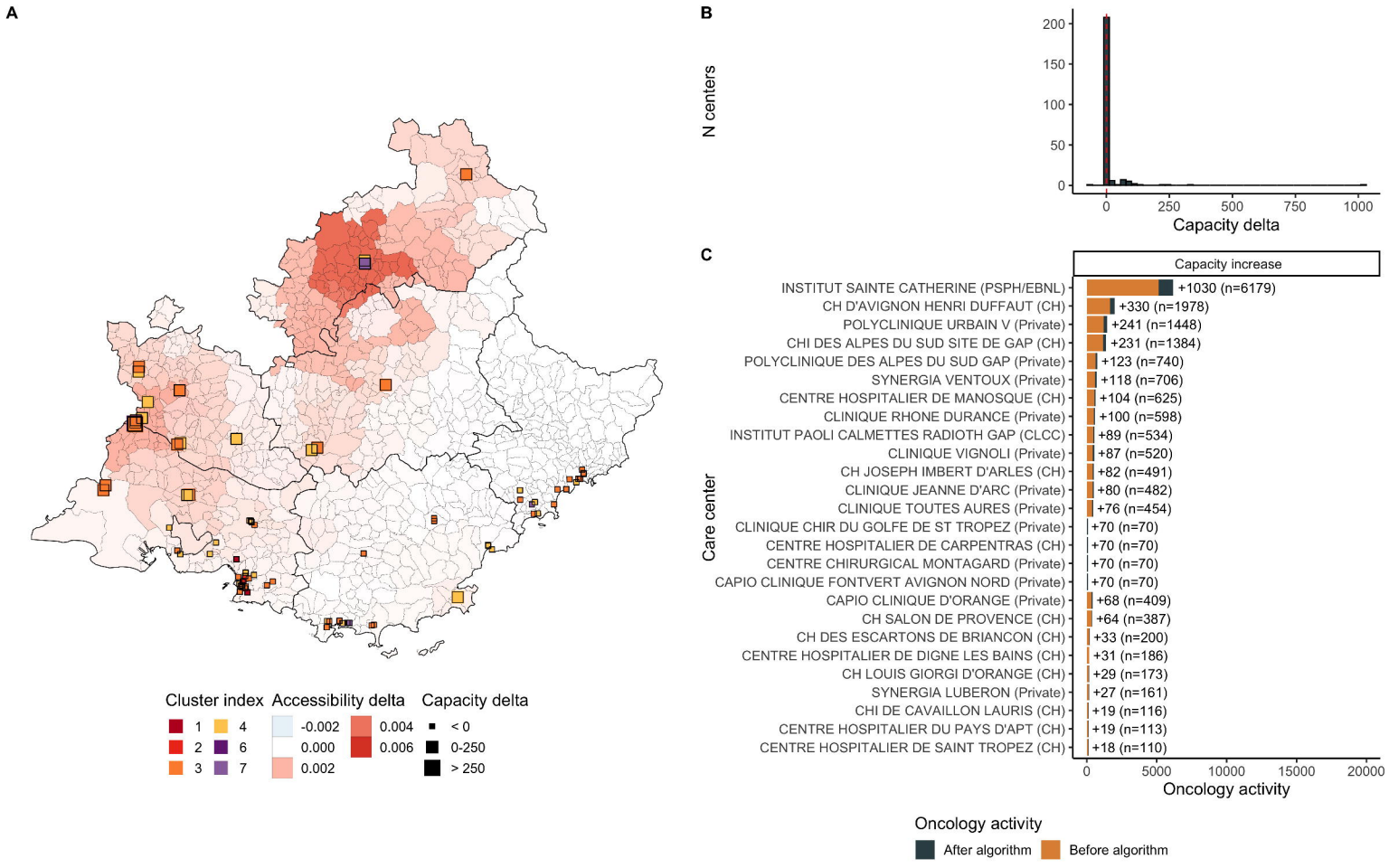

Cluster 8 (n=626)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)


All clusters (n=1662)
Biology
Radiotherapy
Intensive unit
Palliative care
Oncology service
Medication circuit
Cancer surgery
Outpatient surgery
Chemotherapy
Surgery
100 (%)
75 (%)
50 (%)
25 (%)
0 (%)


A
Oncology activity
0
5000
10000
15000
20000
Accessibility quantile
Q1
Q2
Q3
Q4
Q5
NA
Cluster index
1
2
3
4
5
6
7
8
B
Accessibility score
Q5
Q4
Q3
Q2
Q1
33.4 M (51.6%)
13.2 M (20.4%)
8.0 M (12.3%)
6.0 M (9.3%)
4.1 M (6.3%)
0e+00
1e+07
2e+07
3e+07
Population
C
Accessibility score
30%
6%
24%
28%
7%
38%
21%
29%
39%
22%
32%
43%
29%
25%
72%
17%
11%
0
25
50
75
100
Percentage of care centers
Cluster index
D
Top 10
Hérault (Occitanie)
Paris (Ile-de-France)
Hauts-de-Seine (Ile-de-France)
Val-de-Marne (Ile-de-France)
Seine-Saint-Denis (Ile-de-France)
Charente (Nouvelle-Aquitaine)
Bouches-du-Rhône (Provence-Alpes-Côte d'Azur)
Alpes-Maritimes (Provence-Alpes-Côte d'Azur)
Rhône (Auvergne-Rhône-Alpes)
Loire-Atlantique (Pays de la Loire)
Bottom 10
Drôme (Auvergne-Rhône-Alpes)
Gers (Occitanie)
Meuse (Grand Est)
Nièvre (Bourgogne-Franche-Comté)
Haute-Marne (Grand Est)
Eure-et-Loir (Centre-Val de Loire)
Cantal (Auvergne-Rhône-Alpes)
Ariège (Occitanie)
Alpes-de-Haute-Provence (Provence-Alpes-Côte d'Azur)
Lozère (Occitanie)
Departement (region)
0.00
0.02
0.04
Accessibility score


A
B
C
Population density
<30
30-50
50-100
100-200
>200
Accessibility quantile
Q1
Q2
Q3
Q4
Q5
D
E
F
Population density
<30
30-50
50-100
100-200
>200
Average travel time
30
60
90
120
150


A
B
C
Cluster index
1
2
3
4
6
7
8
Oncology activity
0
250
1000
5000
Accessibility quantile
Q1
Q2
Q3
Q4
Q5
Population density (2017)
<30
30-50
50-100
100-200
>200
Average travel time (mins)
30
60
90
120
150
D
Bouches-du-Rhône
Alpes-Maritimes
Var
Hautes-Alpes
Vaucluse
Alpes-de-Haute-Provence
Departement
0.00
0.02
0.04
Accessibility score
E
>200
100-200
50-100
30-50
<30
Density cut
0.00
0.02
0.04
Accessibility score
F
150
100
50
0
Average travel time
90 mins
60 mins
30 mins
0.00
0.02
0.04
Accessibility score


A
Cluster index
1
2
3
4
6
7
Accessibility delta
-0.002
0.000
0.002
0.004
0.006
Capacity delta
< 0
0-250
> 250
B
N centers
200
150
100
50
0
0
250
500
750
1000
Capacity delta
C
Capacity increase
Care center
INSTITUT SAINTE CATHERINE (PSPH/EBNL) +1030 (n=6179)
CH D'AVIGNON HENRI DUFFAUT (CH) +330 (n=1978)
POLYCLINIQUE URBAIN V (Private) +241 (n=1448)
CHI DES ALPES DU SUD SITE DE GAP (CH) +231 (n=1384)
POLYCLINIQUE DES ALPES DU SUD GAP (Private) +123 (n=740)
SYNERGIA VENTOUX (Private) +118 (n=706)
CENTRE HOSPITALIER DE MANOSQUE (CH) +104 (n=625)
CLINIQUE RHONE DURANCE (Private) +100 (n=598)
INSTITUT PAOLI CALMETTES RADIOTH GAP (CLCC) +89 (n=534)
CLINIQUE VIGNOLI (Private) +87 (n=520)
CH JOSEPH IMBERT D'ARLES (CH) +82 (n=491)
CLINIQUE JEANNE D'ARC (Private) +80 (n=482)
CLINIQUE TOUTES AURES (Private) +76 (n=454)
CLINIQUE CHIR DU GOLFE DE ST TROPEZ (Private) +70 (n=70)
CENTRE HOSPITALIER DE CARPENTRAS (CH) +70 (n=70)
CENTRE CHIRURGICAL MONTAGARD (Private) +70 (n=70)
CAPIO CLINIQUE FONTVERT AVIGNON NORD (Private) +70 (n=70)
CAPIO CLINIQUE D'ORANGE (Private) +68 (n=409)
CH SALON DE PROVENCE (CH) +64 (n=387)
CH DES ESCARTONS DE BRIANCON (CH) +33 (n=200)
CENTRE HOSPITALIER DE DIGNE LES BAINS (CH) +31 (n=186)
CH LOUIS GIORGI D'ORANGE (CH) +29 (n=173)
SYNERGIA LUBERON (Private) +27 (n=161)
CHI DE CAVAILLON LAURIS (CH) +19 (n=116)
CENTRE HOSPITALIER DU PAYS D'APT (CH) +19 (n=113)
CENTRE HOSPITALIER DE SAINT TROPEZ (CH) +18 (n=110)
0
5000
10000
15000
20000
Oncology activity
Oncology activity
After algorithm
Before algorithm

SIMCA: SINKHORN MATRIX FACTORIZATION WITH CAPACITY CONSTRAINTS

ERIC DAOUD[1], LUCA GANASSALI[2], ANTOINE BAKER[2] AND MARC LELARGE[2]

[1]*INRIA, DI/ENS, INSTITUT CURIE, PARIS, FRANCE*
[2]*INRIA, DI/ENS, PARIS, FRANCE*

Abstract. For a very broad range of problems, recommendation algorithms have been increasingly used over the past decade. In most of these algorithms, the predictions are built upon user-item affinity scores which are obtained from high-dimensional embeddings of items and users. In more complex scenarios, with geometrical or capacity constraints, prediction based on embeddings may not be sufficient and some additional features should be considered in the design of the algorithm.

In this work, we study the recommendation problem in the setting where affinities between users and items are based both on their embeddings in a latent space and on their geographical distance in their underlying euclidean space (e.g., $\mathbb{R}^2$), together with item capacity constraints. This framework is motivated by some real-world applications, for instance in healthcare: the task is to recommend hospitals to patients based on their location, pathology, and hospital capacities. In these applications, there is somewhat of an asymmetry between users and items: items are viewed as static points, their embeddings, capacities and locations constraining the allocation. Upon the observation of an optimal allocation, user embeddings, items capacities, and their positions in their underlying euclidean space, our aim is to recover item embeddings in the latent space; doing so, we are then able to use this estimate e.g. in order to predict future allocations.

We propose an algorithm (SiMCa) based on matrix factorization enhanced with optimal transport steps to model user-item affinities and learn item embeddings from observed data. We then illustrate and discuss the results of such an approach for hospital recommendation on synthetic data.

1. Introduction

In a very broad range of applications – many of them being led by e-commerce leaders (Amazon [9], Netflix [7]) – recommendation algorithms have been increasingly used over the past decade. These algorithms are capable of showing users a personalized selection of items they may like, based on their interests and user behavior.

Up to now, the predictions are built upon user-item affinity scores (e.g., user/movie ratings) which are obtained from high-dimensional embeddings of items and users. While these approaches work for most e-commerce applications, there are other natural settings in which more attributes should be considered in the recommendation process. For instance, item capacity constraints are of paramount importance in location or route recommendation, where recommending the same item to every user could lead to congestion and significantly deteriorate user experience [1]. Moreover, in the case of location recommendation, travel distance is also a key factor: the user's choice is often the result of a trade-off between affinity and proximity [14]. In the healthcare sector, patients are usually addressed to an hospital by their general practitioner – or by word of mouth. Since the choice of hospital and practitioner may be critical, an important issue is to make sure that patients are routed to the best place possible – namely to a nearby and adapted structure, without capacity saturation.

In this work, we study the recommendation problem in the setting where affinities between users and items are based both on their embeddings embeddings in a latent space and on their geographical distance in their underlying euclidean space (e.g., $\mathbb{R}^2$), together with item

capacity constraints. Upon the observation of an optimal allocation, user embeddings, items capacities, and their positions in the euclidean space, our aim is to recover item embeddings in the latent space; doing so, we are then able to use this estimate e.g. in order to predict future allocations. Our contributions are as follows:

(*i*) we propose an algorithm based on matrix factorization enhanced with optimal transport steps to model user-item affinities and learn item embeddings from observed data;
(*ii*) we then illustrate and discuss the results of such an approach for hospital recommendation on synthetic data.

Paper organization. After reviewing related work, we formally define the problem in mathematical terms, we describe our algorithm for Sinkhorn Matrix Factorization with Capacity Constraints (SiMCa) and give theoretical guarantees on its convergence. We then illustrate our method for the hospital recommendation problem on synthetic data, discussing the results as well as the choice of parameters.

2. Related work

Hospital and practitioner recommendation has already been studied in the literature (see e.g. the survey [12]). However, to the best of our knowledge, no existing method incorporates hospital capacity constraints in the algorithm training. This tends to refer many users to the same hospital, potentially saturating it and degrading the overall care quality.

Matrix factorization [7] is among the most popular collaborative filtering recommendation algorithms. Matrix factorization characterizes every user i and item j by high-dimensional embeddings u_i, v_j, and predict the user-item affinity by the inner product $\langle u_i, v_j \rangle$. This method has already been applied for patient/doctor recommendation [6, 13]. However, regular matrix factorization is usually applied to simple recommendation problems, such as movie recommendation: as already explained before, recommending locations brings new challenges and requires a different approach [14].

Geographical influence has been integrated in the matrix factorization framework to recommend locations or points of interest (POIs) [8]: moreover, the learning algorithm can be adapted by adding a capacity term in the loss function [1].

The Monge-Kantorovitch formulation of the classical Optimal Transport (OT) problem can be rephrased as a linear program that can be computationally slow and unstable in high dimension [2]: this problem is often approximated by adding an entropy regularization term, and easily solved by Sinkhorn-Knopp's algorithm [2]. Another important advantage of this regularization is that the solution of the OT problem becomes differentiable with respect to the parameters, which explains why this step is integrated in many learning algorithms [3, 5, 11].

Most relevant for the present paper is the work from Dupuy, Galichon and Sun [4]. In this study, the authors address the inverse optimal transport problem, that is, given vectors of characteristics $\mathbf{X} \in \mathbb{R}^d$ and $\mathbf{Y} \in \mathbb{R}^{d'}$ and the joint distribution of the optimal matching, the problem of recovering the affinity function of the form $\phi(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T \mathbf{A} \mathbf{Y}$, namely to estimate matrix $\mathbf{A}$. The authors are in the setting where they observe pairs of embeddings $(\mathbf{X}_t, \mathbf{Y}_t)$ together with the optimal *regularized matching* π^* – that is the solution to problem (2) hereafter – and build an estimator of $\mathbf{A}$ with low-rank constraints, the objective being to isolate important characteristics that carry the most important weight in the matching procedure between x and y. We stress the fact that the setting is different in our study: we only observe in our case the embeddings $\mathbf{U}$ of the users and a distance matrix $\mathbf{D}$, function ϕ is known as well as the *pure matching* σ^* – that is the solution of the linear assignment problem (1) hereafter, which differs from π^* – and the aim is to infer item embeddings $\mathbf{V}$. In other words, we do not seek to reconstruct the affinity matrix, but for the learning of items' positions in the user's embeddings space, these positions acting as reference points, upon which prediction of future allocations can be made. Another difference is that the number of

items is typically very small compared to the number of users, which justifies that the items are considered static: we also incorporate *capacity constraints* on the allocation problem.

3. Problem definition

A model for latent and geographical affinity. The setting of the problem is as follows. Consider n *users* $x_1, \dots, x_n$ embedded in a latent space $\mathcal{X}$ identified to $\mathbb{R}^d$, with embeddings given by $\mathbf{U}_1, \dots, \mathbf{U}_n$. Also consider m *items* $y_1, \dots, y_m$ embedded in $\mathcal{X}$ with embeddings $\mathbf{V}_1, \dots, \mathbf{V}_m$, with $m \leq n$. To each user x_i we assign a single item y_j, according to an *affinity matrix* $\mathbf{M} \in \mathbb{R}^{n \times m}$ given by

$$\mathbf{M}_{i,j} := \Phi(\mathbf{U}_i, \mathbf{V}_j, \mathbf{D}_{i,j}),$$

where $\mathbf{D} \in \mathbb{R}^{n \times m}$ is known and may be thought of e.g. as a geographical distance matrix between users and items in the underlying euclidean space, say $\mathbb{R}^2$ (we stress the fact that this space is *not* the embedding space $\mathcal{X}$). We will denote $\mathbf{M} = \Phi(\mathbf{U}, \mathbf{V}, \mathbf{D})$ in the sequel.

We also work under the following constraints: each item $y_j, j \in [m]$ can be assigned to at most $\mathbf{C}_j$ users. Where $\mathbf{C} = (\mathbf{C}_1, \dots, \mathbf{C}_m)$ is *capacity* vector. The *total capacity* is defined by

$$s(\mathbf{C}) := \sum_{j \in [m]} \mathbf{C}_j,$$

and we will assume $s(\mathbf{C}) = n$. We define

$$\Sigma(n, m, \mathbf{C}) := \left\{ \sigma \in \{0,1\}^{n \times m},\ \sigma \mathbf{1}_m = \mathbf{1}_n, \sigma^T \mathbf{1}_n = \mathbf{C} \right\}.$$

In the sequel, σ will denote both the assignment and its corresponding matrix representation. The optimal assignment σ^* is given by

$$\sigma^*(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{C}) := \underset{\sigma \in \Sigma(n,m,\mathbf{C})}{\arg\max} \operatorname{Tr}\left(\sigma^T \mathbf{M}\right), \tag{1}$$

Note that problem (1) is an instance of the *Linear Assignment problem* (LAP).

Goal. Assume that we are given the user embeddings $\mathbf{U}$, the distance matrix $\mathbf{D}$, the capacities $\mathbf{C}$ and the optimal assignment $\sigma^* \in \Sigma(n, m, \mathbf{C})$. The goal is to learn the item embeddings $\mathbf{V}$.

Loss metrics, regularization and relaxation. We will evaluate the performance of a proposed estimate $\widehat{\mathbf{V}}$ of $\mathbf{V}$ through the assignment $\widehat{\pi}$ obtained with $\widehat{\mathbf{V}}$. To compare $\widehat{\pi}$ with σ^*, we use the usual *cross entropy loss* defined by

$$H(\sigma^*, \widehat{\pi}) := -\sum_{i \in [n]} \log \widehat{\pi}_{i, \sigma^*(i)} = -\operatorname{Tr}\left((\sigma^*)^T (\log \widehat{\pi})\right).$$

As stated before, from a learning perspective, a main issue is that the solution to problem (1) is not differentiable w.r.t. $\mathbf{V}$, the variable of interest. This issue is solved by a relaxation/regularization procedure [2]:

- since the objective function is linear, we first consider the classical relaxation of (1) on the polytope of the convex hull of $\Sigma(n, m, \mathbf{C})$, namely on

$$\Pi(n, m, \mathbf{C}) := \left\{ \pi \in [0,1]^{n \times m},\ \pi \mathbf{1}_m = \mathbf{1}_n, \pi^T \mathbf{1}_n = \mathbf{C} \right\}.$$

- moreover, we regularize the objective function in order to perform (automatic) differentiation: this is made possible by the classical entropy regularization in optimal transport.

For a small regularization parameter $\varepsilon > 0$, the problem then becomes

$$\pi^*_\varepsilon(\mathbf{U}, \mathbf{V}, \mathbf{D}, \mathbf{C}) := \underset{\pi \in \Pi(n,m,\mathbf{C})}{\arg\max} \left[\operatorname{Tr}\left(\pi^T \mathbf{M}\right) + \varepsilon H(\pi)\right], \tag{2}$$

where

$$H(\pi) := -\sum_{1 \leq i,j \leq n} \pi_{i,j}(\log \pi_{i,j} - 1). \tag{3}$$

It is known in the literature [2] that the solution π_ε^* to the convex optimization problem (2) can be easily computed with Sinkhorn-Knopp's algorithm, and has the following form:

$$(\pi_\varepsilon^*)_{i,j} = a_i \exp\left(\frac{1}{\varepsilon}\mathbf{M}_{i,j}\right) b_j, \tag{4}$$

where a and b are vectors of $\mathbb{R}_+^n$ and $\mathbb{R}_+^m$. Note that we are back to our initial problem (1) when $\varepsilon = 0$.

SiMCa Algorithm. With this new formulation (2), we are now able to design an optimization scheme for our learning problem. In our setting the users embeddings $\mathbf{U}$, the distance matrix $\mathbf{D}$ and the capacities $\mathbf{C}$ are known, only the items embeddings $\mathbf{V}$ are learned. The overall procedure is summarized in Algorithm 1. Given the current estimate $\mathbf{V}_t$ at iteration t, we compute the solution $\pi_\varepsilon^*(\mathbf{V}_t)$ to problem (2), which in turn is used to compute the gradient in $\mathbf{V}_t$ of the following loss

$$\text{loss}(\mathbf{V}_t) := H\left(\sigma^*, \pi_\varepsilon^*(\mathbf{V}_t)\right) \tag{5}$$

to update our estimate of $\mathbf{V}$ through a gradient step. The gradient in $\mathbf{V}$ has actually a simple analytical expression:

Lemma 1. *We have*

$$\nabla_{\mathbf{V}}\text{loss}(\mathbf{V}) = \frac{1}{\varepsilon} \sum_{1\le i,j\le n} (\pi_\varepsilon^*(\mathbf{V}) - \sigma^*)_{i,j} \nabla_{\mathbf{V}}\mathbf{M}_{i,j}\,. \tag{6}$$

Proof. A very similar expression for the gradient is derived for the maximum likelihood in [4]. We straightforwardly adapt their derivation to the cross entropy loss (5). Let us denote

$$V_\varepsilon(\mathbf{M}) = \max_{\pi\in\Pi(n,m,\mathbf{C})} \left[\text{Tr}\left(\pi^T\mathbf{M}\right) + \varepsilon H(\pi)\right] \tag{7}$$

the optimal value of the regularized OT problem (2). As well-known in the OT literature, see Proposition 9.2 of [10], its gradient with respect to the affinity matrix M is given by the optimal coupling

$$\frac{\partial}{\partial \mathbf{M}_{i,j}} V_\varepsilon(\mathbf{M}) = (\pi_\varepsilon^*)_{i,j}. \tag{8}$$

Our cross-entropy loss (5) is directly related to the optimal value $V_\varepsilon(\mathbf{M})$:

$$\begin{aligned}
\text{loss} = H\left(\sigma^*, \pi_\varepsilon^*\right) &= -\sum_{i,j} \sigma_{i,j}^* \ln(\pi_\varepsilon^*)_{i,j} \\
&\overset{1}{=} -\sum_{i,j} \sigma_{i,j}^* (\tfrac{1}{\varepsilon}\mathbf{M}_{i,j} + \ln a_i + \ln b_j) \\
&\overset{2}{=} -\sum_{i,j} \sigma_{i,j}^* \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - \sum_{i,j} (\pi_\varepsilon^*)_{i,j} (\ln a_i + \ln b_j) \\
&\overset{3}{=} -\sum_{i,j} \sigma_{i,j}^* \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - \sum_{i,j} (\pi_\varepsilon^*)_{i,j} (\ln(\pi_\varepsilon^*)_{i,j} - \tfrac{1}{\varepsilon}\mathbf{M}_{i,j}) \\
&\overset{4}{=} -\sum_{i,j} \sigma_{i,j}^* \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} - s(\mathbf{C}) \\
&\quad - \sum_{i,j} (\pi_\varepsilon^*)_{i,j} (\ln(\pi_\varepsilon^*)_{i,j} - 1) + \sum_{i,j} (\pi_\varepsilon^*)_{i,j} \tfrac{1}{\varepsilon}\mathbf{M}_{i,j} \\
&\overset{5}{=} -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[\text{Tr}(\pi_\varepsilon^{*T}\mathbf{M}) + \varepsilon H(\pi_\varepsilon^*) - \text{Tr}(\sigma^{*T}\mathbf{M})] \\
&\overset{6}{=} -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[V_\varepsilon(\mathbf{M}) - \text{Tr}(\sigma^{*T}\mathbf{M})].
\end{aligned}$$

The first and third equalities follow from (4), the second and fourth from $\sigma^*, \pi_\varepsilon^* \in \Pi(n, m, \mathbf{C})$, the fifth from the definition (3) of $H(\pi)$ and the sixth from the definition (7) of $V_\varepsilon(\mathbf{M})$. Then differentiating with respect to $\mathbf{V}$ leads to (6) by the chain rule and (8). □

The performance of our method is guaranteed by the following:

Algorithm 1 Sinkhorn Matrix Factorization with Capacity Constraints (SiMCa)

Input: $\mathbf{U}, \mathbf{D}, \mathbf{C}, \sigma^*$
For $t = 1$ to T:

1. Compute the affinity matrix $\mathbf{M}_{t-1} = \Phi(\mathbf{U}, \mathbf{V}_{t-1}, \mathbf{D})$.
2. Compute the solution to the optimization problem (2):
$$\pi^*_\varepsilon(\mathbf{V}_{t-1}) := \underset{\pi \in \Pi(n,m,\mathbf{C})}{\arg\max} \left[\mathrm{Tr}\left(\pi^T \mathbf{M}_{t-1}\right) + \varepsilon H(\pi)\right].$$
3. Compute the gradient $\nabla \mathrm{loss}(\mathbf{V}_{t-1})$ with equation (6).
4. Perform a gradient step $\mathbf{V}_t = \mathbf{V}_{t-1} - \eta \nabla \mathrm{loss}(\mathbf{V}_{t-1})$.

return $\mathbf{V}_T$

Lemma 2. *Assume that $v \mapsto \Phi(u, v, d)$ is linear. Then the loss function* (5) *is convex in* $\mathbf{V}$ *and the output of SiMCa Algorithm (Algo. 1) converges to*

$$\underset{\mathbf{V}}{\arg\min}\, H\left(\sigma^*, \pi^*_\varepsilon(\mathbf{V})\right).$$

Proof. The proof of Lemma 1 shows that

$$\mathrm{loss}(V) = -s(\mathbf{C}) + \tfrac{1}{\varepsilon}[V_\varepsilon(\mathbf{M}) - \mathrm{Tr}(\sigma^{*T}\mathbf{M})].$$

Since $V \mapsto \Phi(\mathbf{U}, V, \mathbf{D})$ is linear, $V \mapsto V_\varepsilon(\mathbf{M})$, as defined in (7) is convex as a maximum of convex functions. By assumption, $V \mapsto \mathrm{Tr}(\sigma^{*T}\mathbf{M})$ is linear, thus $V \mapsto \mathrm{loss}(V)$ is convex. □

4. Illustration for the hospital recommendation problem

We now describe an illustration of our method for the hospital recommendation problem. Since very few open datasets are available for this problem, we trained our algorithm on synthetic data.

Dataset generation. The dataset is generated as follows:

- **Features in the embedding (latent) space:** we sample $n + m$ points from a Gaussian mixture model with k clusters. We set these points as either users ($\mathbf{U}_i$) or items ($\mathbf{V}_i$), and considered that each cluster must contain at least one item: we are thus left with n users and m items, spread between k clusters. Users and items in the same cluster are considered similar. We then normalized both users and items features, so that all embeddings $\mathbf{U}_i$ and $\mathbf{V}_j$ lie on the unit sphere. Note that the users and items sampling is done independently of items capacities.
- **Distance in the underlying euclidean space:** to sample the distance matrix $\mathbf{D}$ between users and items, we sample all the positions randomly on a circle, and computed the great-circle distance (i.e. spherical distance) between every users i and items j. We finally normalize the distance matrix by its overall mean.
- **Capacities** we sampled m values from a Dirichlet Distribution, corresponding to the probabilities that users are assigned to the m items. We converted these probabilities into capacities $\mathbf{C}_j$ by multiplying them with the number of users n. We then added some extra spots to each item.

Affinity matrix. In our case, the affinity matrix $\mathbf{M} = \Phi(\mathbf{U}, \mathbf{V}, \mathbf{D})$ is defined as follows:

$$\mathbf{M}_{i,j} = \Phi(\mathbf{U}_i, \mathbf{V}_j, \mathbf{D}_{i,j}) = (1 - \alpha)\mathbf{U}_i^T \mathbf{V}_j - \alpha \mathbf{D}_{i,j}. \tag{9}$$

The α coefficient measures the trade-off between affinity and proximity.

We then solve the Linear Assignment Problem (1) to compute the pure matching σ^*.

Noise. Noise is added to the original dataset in two different ways. The first method is to modify the allocations of random users in σ^*, the noise ratio being defined as the percentage of modified allocations[1]. The second method consists in perturbating $\mathbf{U}$ as follows:

$$\widetilde{\mathbf{U}} := \sqrt{1-\rho^2}\mathbf{U} + \rho\mathbf{Z},$$

where $\mathbf{Z}$ is a matrix with i.i.d. standard Gaussian entries, and ρ is the noise ratio.

Learning the embeddings. Given $\mathbf{U}$, $\mathbf{D}$, $\mathbf{C}$, σ^*, α and ε, we compute an estimate $\widehat{\mathbf{V}}$ of the item embeddings with SiMCa Algorithm (Algo. 1). Comparing $\widehat{\mathbf{V}}$ with $\mathbf{V}$ gives a first measure of the training performance.

Recovering the pure matching. Then, using $\widetilde{\mathbf{U}}$ (the noisy version of $\mathbf{U}$), $\widehat{\mathbf{V}}$ (the estimated $\mathbf{V}$), $\mathbf{D}$, α and ε, we compute the solution $\widehat{\pi}^*_\varepsilon$ to problem (2). Solving the linear assignment problem (LAP) on matrix $\widehat{\pi}^*_\varepsilon$, we compute a pure matching $\widehat{\sigma}^*$, which we can next compare to the original σ^*, giving a second measure of the training performance.

5. Results

Parameters. We generated a toy dataset with the following parameters: $n = 1000$ users; $m = 3$ items; $d = 2$ latent features; $k = 3$ clusters; $\alpha = 0.3$. The items capacities were 257, 417 and 356. Figure 1 shows the generated users and items in both the embeddings (latent) space and their underlying euclidean space.

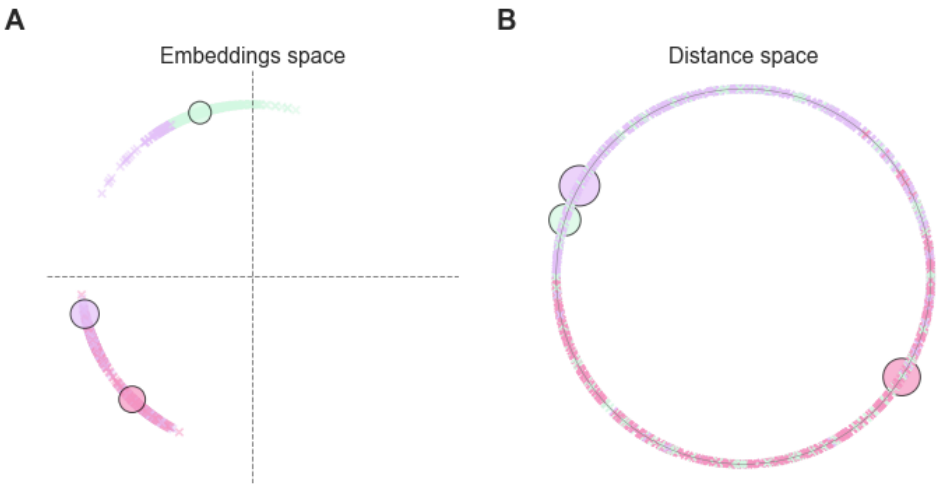


Figure 1. Generated dataset. Users are displayed as crosses and items as circles, sized proportionally to their capacities. Users are colored accordingly to the item they have been allocated to. Plot (A) displays users and items in their shared embeddings (latent) space; where plot (B) displays them in their underlying euclidean space.

Training results. We trained our model with $\varepsilon = 0.1$ entropy regularization and 10 iterations in Sinkhorn-Knopp's algorithm to output $\widehat{\mathbf{V}}$. As mentioned earlier we compute a solution to the LAP on matrix $\widehat{\pi}^*_\varepsilon$ to output the estimated allocation $\widehat{\sigma}^*$. The model was trained with Adam optimizer, with a 0.01 learning rate and 400 epochs. For the measures of performance, we used the F1 score, for measuring how well the allocations are reproduced, and the mean euclidean distance between learned embeddings $\widehat{\mathbf{V}}$ and the ground truth $\mathbf{V}$. The training results are displayed on Figure 2.

[1]to make sure that the capacities constraints on the items still hold, we must swap *pairs of users*: for a given allocation to modify, we pick another user randomly and swap their allocations.

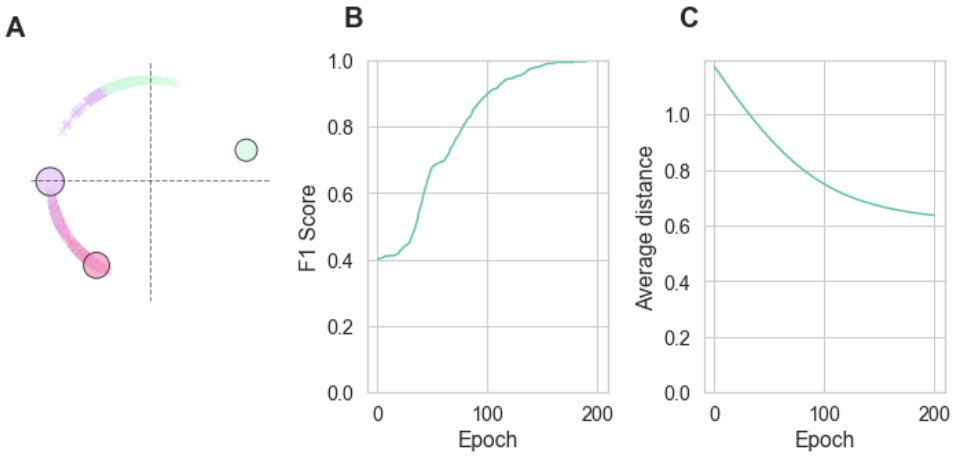


FIGURE 2. Training results. We can see that the model achieves good performances to learn the item embeddings (A), and recovers the allocation with a close to 1 F1 score (B). The average distance between the learned embeddings and ground truth decreases during training (C).

Influence of entropy regularization. We investigate the influence of the entropy regularization parameter ε on the model performance. We let ε vary between 0.05 and 2, with the same dataset and the same model parameters[2]. We ran 5 training for every value of ε. As shown on Figure 3, the training performance worsens when ε increases.

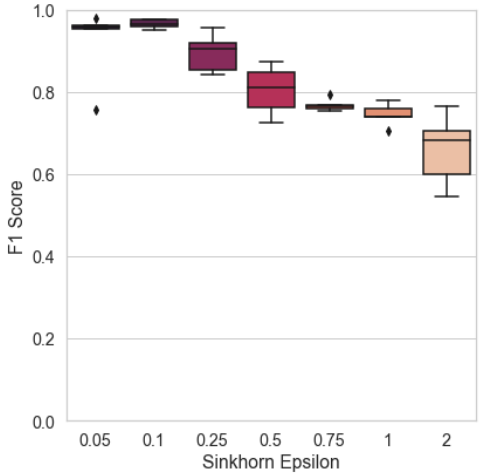


FIGURE 3. Performance as a function of ε. Increasing ε leads to lower F1 scores.

Influence of noise. We study the influence of noise, either by swapping allocations or adding Gaussian noise to the used embeddings, as described in the previous section. Unsurprisingly, as shown in Figure 4, the training performance is decreasing with the noise ratio.

Learning both users and items embeddings. We also studied the case where the users' embeddings are not known, and must be learned jointly with the items' embeddings from the observed allocations. In this case, we initialized the items' and users' embeddings similarly. We managed to retrieve the observed allocation as illustrated on Figure 5. However, the

[2]due to numerical instability, the algorithm could not train properly above $\varepsilon = 0.05$.

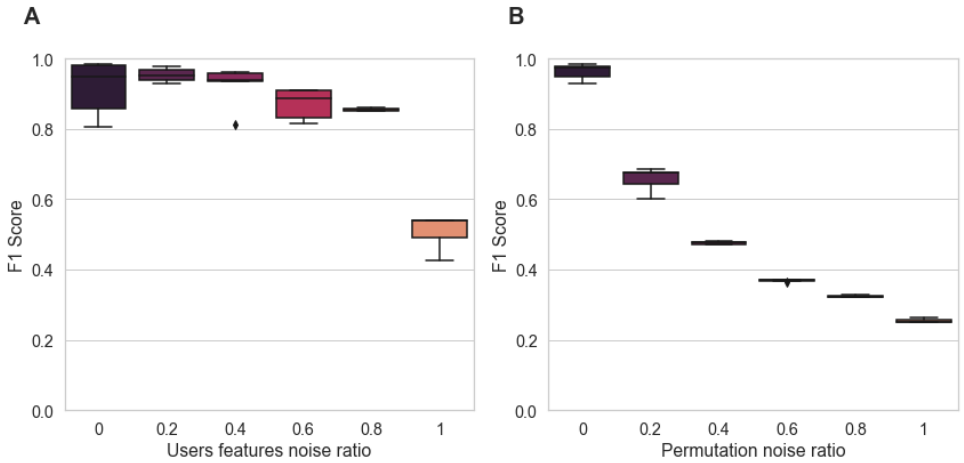


FIGURE 4. Influence of adding Gaussian noise (A) and swapping allocations (B) on training performance. F1 score decreases as noise increases.

average distance between learned items embeddings and ground truth does not decrease during training, meaning that the model learned its own interpretation of the users' and items' representations to satisfy the observed mapping.

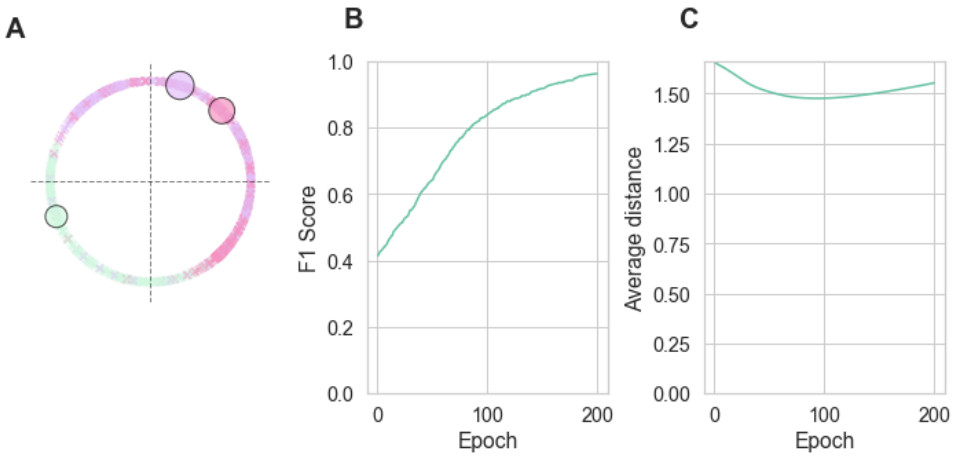


FIGURE 5. Learning both users and items embeddings simultaneously. The learned embeddings are shown on (A). The model retrieves the observed allocation (B). However, the average distance between learned items embeddings and ground truth does not decrease during training (C).

6. CONCLUSIONS

In this work, we introduced SiMCa, an algorithm based on matrix factorization and optimal transport to model user-item affinities and learn item embeddings from observed data. SiMCa can be used in recommendation problems where allocations between users and items are based on: their affinity in a latent space; a geographical distance in their underlying euclidean space; capacity constraints on the items. We illustrated our method for the hospital recommendation task; however, we believe that there are many other problems for which SiMCa algorithm may be useful.

Acknowledgments

This work was partially supported by the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR19-P3IA-0001 (PRAIRIE 3IA Institute).

References

[1] Konstantina Christakopoulou, Jaya Kawale, and Arindam Banerjee. Recommendation with Capacity Constraints. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1439–1448, Singapore Singapore, November 2017. ACM.

[2] Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transportation Distances. *arXiv:1306.0895 [stat]*, June 2013. arXiv: 1306.0895.

[3] Marco Cuturi and Mathieu Blondel. Soft-DTW: a Differentiable Loss Function for Time-Series. *arXiv:1703.01541 [stat]*, February 2018. arXiv: 1703.01541.

[4] Arnaud Dupuy, Alfred Galichon, and Yifei Sun. Estimating matching affinity matrix under low-rank constraints. *arXiv:1612.09585 [stat]*, December 2016. arXiv: 1612.09585.

[5] Aude Genevay, Gabriel Peyré, and Marco Cuturi. Learning Generative Models with Sinkhorn Divergences. *arXiv:1706.00292 [stat]*, October 2017. arXiv: 1706.00292.

[6] Qiwei Han, Mengxin Ji, Inigo Martinez de Rituerto de Troya, Manas Gaur, and Leid Zejnilovic. A Hybrid Recommender System for Patient-Doctor Matchmaking in Primary Care. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 481–490, October 2018.

[7] Yehuda Koren, Robert Bell, and Chris Volinsky. Matrix Factorization Techniques for Recommender Systems. *Computer*, 42(8):30–37, August 2009. Conference Name: Computer.

[8] Xutao Li, Gao Cong, Xiao-Li Li, Tuan-Anh Nguyen Pham, and Shonali Krishnaswamy. Rank-GeoFM: A Ranking based Geographical Factorization Method for Point of Interest Recommendation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '15, pages 433–442, New York, NY, USA, August 2015. Association for Computing Machinery.

[9] G. Linden, B. Smith, and J. York. Amazon.com recommendations: item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76–80, January 2003. Conference Name: IEEE Internet Computing.

[10] Gabriel Peyré and Marco Cuturi. Computational Optimal Transport. *arXiv:1803.00567 [stat]*, March 2020. arXiv: 1803.00567.

[11] Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Sinkhorn Label Allocation: Semi-Supervised Classification via Annealed Self-Training. *arXiv:2102.08622 [cs, stat]*, June 2021. arXiv: 2102.08622.

[12] Thi Ngoc Trang Tran, Alexander Felfernig, Christoph Trattner, and Andreas Holzinger. Recommender systems in the healthcare domain: state-of-the-art and research issues. *Journal of Intelligent Information Systems*, 57(1):171–201, August 2021.

[13] Yin Zhang, Min Chen, Dijiang Huang, Di Wu, and Yong Li. iDoctor: Personalized and professionalized medical recommendations based on hybrid matrix factorization. *Future Generation Computer Systems*, 66:30–35, January 2017.

[14] Shenglin Zhao, Irwin King, and Michael R. Lyu. A Survey of Point-of-interest Recommendation in Location-based Social Networks. *arXiv:1607.00647 [cs]*, July 2016. arXiv: 1607.00647.