Charles University in Prague

Faculty of Social Sciences
Institute of Economic Studies

MASTER'S THESIS

Czech Allergology and Clinical Immunology: Utilization and Provision in Space

Author: **Bc. Lucie Blechová**

Supervisor: **Mgr. Michal Paulus**

Academic Year: **2015/2016**

Declaration of Authorship

The author hereby declares that she compiled this thesis independently, using only the listed resources and literature, and the thesis has not been used to obtain a different or the same degree.

The author grants to Charles University permission to reproduce and to distribute copies of this thesis document in whole or in part.

Prague, January 4, 2016

Signature

Acknowledgments

I am especially grateful to my supervisor Mgr. Michal Paulus for his priceless advice throughout the whole process of writing this thesis. I am also very grateful to PhDr. Lenka Šťastná Ph.D. who consulted spatial models with me. My thanks belong also to PhDr. Ladislav Krištoufek Ph.D. for recommendations in topics of general econometrics.

I would also like to express my gratitude to the Ministry of Health of the Czech Republic and to the Institute of Health Information and Statistics of the Czech Republic for providing me with data which enabled me to study this topic.

Abstract

The present thesis focuses on Czech allergology and clinical immunology by studying this branch of health care in a geographic variation framework while using methods of spatial econometrics. This has been the first work with such focus. District-level data on care provision and utilization in 2012 are used. It is found that there exist geographical differences between provision and utilization and that the geographical distribution of allergists and clinical immunologists does not correspond to population's needs. Care utilization is modeled using a spatial autoregressive model specification. Based on this model, it is concluded that a shortage of physicians in the majority of districts actually limits care utilization. Also based on the utilization model, there is a discussion about the potential need for policy coordination. Care provision cannot be modeled using explanatory variables that are available, therefore, future data collection is necessary. However, it was found that variables influencing the need for care by patients do not influence care provision per physician.

JEL Classification	I11, I14, I19
Keywords	geographic variation, Czech, allergology, clinical immunology, ACI, spatial model

Author's e-mail	18702440@fsv.cuni.cz
Supervisor's e-mail	michal.paulus@fsv.cuni.cz

Abstrakt

Tato práce se zaměřuje na českou alergologii a klinickou imunologii a studuje toto odvětví v rámci geografické rozdílnosti za použití metod prostorové ekonometrie. Jedná se o první práci s tímto zaměřením. Pro poskytování a čerpání péče jsou použita data z roku 2012 na úrovni okresů. Je zjištěna geografická rozdílnost mezi čerpáním a poskytováním péče a také je zjištěno, že geografické rozložení alergologů a klinických imunologů neodpovídá potřebám populace. Čerpání péče je modelováno pomocí prostorového autoregresivního modelu. Na základě tohoto modelu je zjištěno, že nedostatek lékařů ve většině okresů omezuje čerpání péče. Také je na jeho základě diskutována možnost nutnosti koordinace politik. Poskytování péče nelze modelovat za použití v současnosti dostupných vysvětlujících proměnných, a tedy je nutné se v budoucnosti zaměřit na sběr dat. Bylo ale zjištěno, že proměnné ovlivňující potřebu péče pacientů nemají vliv na poskytování péče na lékaře.

Klasifikace	I11, I14, I19
Klíčová slova	geografická rozdílnost, Česko, alergologie, klinická imunologie, ALG, prostorový model

E-mail autora	18702440@fsv.cuni.cz
E-mail vedoucího práce	michal.paulus@fsv.cuni.cz

Contents

List of Tables

List of Figures

Acronyms

ACI Allergology and Clinical Immunology

AIC Akaike Information Criterion

CR Czech Republic

EU European Union

IHIS Institute of Health Information and Statistics of the Czech Republic

LAU Local Administrative Units

LHS List of Health Services

LM Lagrange Multiplier

LR Likelihood Ratio

MH Ministry of Health of the Czech Republic

MLE Maximum Likelihood Estimation

OLS Ordinary Least Squares

PC Principal Component

PCA Principal Component Analysis

SAR Spatial Autoregressive Model

SDM Spatial Durbin Model

SEM Spatial Error Model

SHI Statutory Health Insurance

Master's Thesis Proposal

Author:	Bc. Lucie Blechová
Supervisor:	Mgr. Michal Paulus
Defense Planned:	January 2016

Proposed Topic:

Medical Practice Variation in Czech Allergology

Motivation:

It has previously been found that medical practice varies geographically and it is therefore likely to affect the provision of health care within the Czech Republic. The ambition of this thesis is to explore one particular part of Czech health care – allergology.

The study of medical practice variation is underlined by the thought that in many cases this variation is not caused by differences in needs or preferences of patients (OECD, 2014). Hence the "undesirable" variation can cause higher medical costs as the treatment provided might not be necessary or effective for the patient. It is thus crucial to study this variation sufficiently in order to guarantee equity and efficiency of health care.

According to de Monchy et al. (2013), about half of Europeans tend to develop allergic reactions and allergic diseases are among those most frequent in the Western World. Thus allergology is a part of health care whose quality has impact on a large part of population. One can therefore expect that variations in this field have a large impact on the overall quality of the Czech health care.

Hypotheses:

1. Hypothesis #1: The care provided in allergology differs across districts in the Czech Republic. According to OECD (2014), many other areas of expertise suffer from medical practice variation. The aim is to test that Czech allergology is not an exception.

2. Hypothesis #2: The volume of care provided depends on the population density in a given district as there tend to be more physicians per capita in larger cities (Czech Statistical Office, 2013).

3. Hypothesis #3: The volume of care provided depends on the environment in a given district. This leads to a higher provision of care in larger cities because these are usually more polluted (Czech Hydrometheorological Institute, 2014).

Methodology:

I am going to use production data provided by the Ministry of Health of the Czech Republic. The data set includes data from the year 2012 which are sorted according to regions and districts of the Czech Republic. The data are related to several areas of

expertise from which I chose allergology. I will then study the medical practice variation for DRG bases in this major diagnostic category. I will use the sum of casemix for each DRG base and model its variation across districts.

The explanatory variables of the model will be chosen within three main categories: socio-economic indicators, health care indicators, and environmental statistics. Socio-economic indicators are used to describe economic well-being in a given district, the underlying idea being that richer people can afford higher quality food and therefore prevent incidence of allergy (ČPZP, 2009). Such indicators available on the district level for the Czech Republic are the unemployment rate and average wage. Health care indicators describe the supply side and they are e.g. number of physicians per 1000 inhabitants of a given district, or if data is available number of allergists per 1000 inhabitants of that district. When it comes to allergology, environmental factors are very important as e.g. dust particles are among the most common allergens (Bartra et al., 2007). The indicators used will be investment in environment protection (against air pollution, water pollution, and waste handling), the size of protected areas within a district, and also the population density for each district as urban areas tend to be more polluted than rural areas. I will use data by the Czech Statistical Office.

The proposed econometric approach for the model is the Ordinary Least Squares (OLS). The method is standardly used in health economics (e.g. Shaw, Horrace & Vogel, 2005). Naturally, all of the assumptions of this method will be properly tested to conclude whether this method is truly the appropriate one for the given dataset. In case of violation of assumptions, the methodology will be modified accordingly.

Hypothesis #1: Analysis of the data for each DRG base and its variation across districts. Only description of the variation, not focusing on its causes (as in OECD, 2014).

Hypothesis #2: The population density is positively correlated with the dependent variable.

Hypothesis #3: Investment in environment protection and the size of protected areas are both negatively correlated with the dependent variable.

Expected Contribution:

The contribution of this work is empirical. Its main aim is to provide an analysis of differences in care in allergology within the Czech Republic. Some other areas of Czech health care have been studied in the medical practice variation framework (OECD, 2014), yet allergology has never been one of them. Thus this work will explore a new area of expertise using tools that were not yet used to study medical practice variation in the Czech Republic.

I hope to make a good use of the data provided to me by creating an analysis that could be used in practice in order to find those districts of the Czech Republic where the quality of care in the particular field of practice is low and also to find common characteristics of such districts in order to improve the quality of care in the future. I hope that my findings could help as a source of information that could be used to explore and improve the equity of health care provision in the Czech Republic.

Outline:

1. Introduction: description of allergology, motivation underlying my interest in this field, medical practice variation description
2. Literature review: studies of allergology in the Czech Republic and in other countries, studies of medical practice variation of other parts of Czech health care, studies of medical practice variation in other countries (hoping to find one on allergology abroad)
3. Data description: detailed description of the provided dataset including its sorting and filtering for the purposes of this thesis
4. The analysis: modeling the variation of the casemix among districts, testing hypotheses
5. Results: providing and commenting the results of my findings and comparing them to results of other studies of medical practice variation
6. Conclusion

Core Bibliography:

BARTRA, J., MULLOL, J., DEL CUVILLO, A., DÁVILLA, I., FERRER, M., JÁUREGUI, I., MONTORO, J., SASTRE, J. & VALERO, A. (2007) Air Pollution and Allergens. Journal of Investigational Allergology and Clinical Immunology. [Online] vol. 17, p. 3-8. Available from: http://www.jiaci.org/issues/vol17s2/2.pdf [Accessed: 2015-01-30].

CORALLO, A. N., CROXFORD, R., GOODMAN, D. C., BRYAN, E. L., SRIVASTAVA, D. & STUKEL, T. A. (2014) A systematic review of medical practice variation in OECD countries. Health Policy. [Online] vol. 114, issue 1, p. 5-14. Available from: http://www.healthpolicyjrnl.com/article/S0168-8510(13)00215-7/abstract [Accessed 2015-01-22].

ČPZP. (2009) Alergii často způsobuje oblíbené jídlo. [online]. Available from: http://www.cpzp.cz/clanek/347-0-Alergii-casto-zpusobuje-oblibene-jidlo.html [Accessed: 2015-01-30].

CZECH HYDROMETEOROLOGICAL INSTITUTE. (2014) Graphic Yearbook 2013. [Online]. Available from: http://portal.chmi.cz/files/portal/docs/uoco/isko/grafroc/13groc/gr13e/Obsah_GB.html [Accessed 2015-02-01].

CZECH STATISTICAL OFFICE. (2013) Public Database: Health care security (ZDR5012PU_OK). [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabparam.jsp?voa=tabulka&cislotab=ZDR5012PU_OK&vo=null [Accessed 2015-01-30].

DE MONCHY, J. G., DEMOLY, P., AKDIS, C. A., CARDONA, V., PAPADOPOULOS, N. G., SCHMID-GRENDELMEIER, P. & GAYRAUD, J. (2013) Allergology in Europe, the blueprint. Allergy. [Online], vol. 68, issue 10, p. 1211-1218. Available from: http://onlinelibrary.wiley.com/doi/10.1111/all.12225/abstract [Accessed 2015-01-22].

OECD. (2014) Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance?. OECD Health Policy Studies. [Online] OECD Publishing. Available from: http://www.oecd-ilibrary.org/social-issues-migration-health/geographic-variations-in-health-care_9789264216594-en [Accessed 2015-01-13].

SHAW, J. W., HORRACE, W. C. & VOGEL, R. J. (2005) The Determinants of Life Expectancy: An Analysis of the OECD Health Data. Southern Economic Journal. [Online], vol. 71, p. 768-783. Available from: http://faculty.maxwell.syr.edu/whorrace/publications/SEJ2005.pdf [Accessed 2015-03-09].

ÚZIS ČR. (2013) Činnost zdravotnických zařízení ve vybraných oborech léčebně preventivní péče 2012. [Online] Praha: Ústav zdravotnických informací a statistiky ČR. Available from: http://www.uzis.cz/katalog/zdravotnicka-statistika/cinnost-zdravotnickych-zarizeni-ve-vybranych-

oborech [Accessed 2015-01-22].

WARNER, J. O., KALINER, M. A., CRISCI, C. D., DEL GIACCO, S., FREW, A. J., GH, L., MASPERO, J., MOON, H., NAKAGAWA, T., POTTER, P. C., ROSENWASSER, L. J., SINGH, A. B., VALOVIRTA, E. & VAN CAUWENBERGE, P. (2006) Allergy Practice Worldwide: A Report by the World Allergy Organization Specialty and Training Council. *Allergy and Clinical Immunology International - World Allergy Organization Journal.* [Online], vol. 18, p. 4-10. Available from: http://www.researchgate.net/publication/7383553_Allergy_practice_worldwide_a_report_by_the_World_Allergy_Organization_Specialty_and_Training_Council [Accessed 2015-01-30].

WESTERT, G.P., GROENEWEGEN, P.P., BOSHUIZEN, H.C., SPREEUWENBERG, P.M.M. & STEULTJENS, M.P.M. (2003) Medical practice variations in hospital care; time trends of a spatial phenomenon. Health & Place. [Online] vol. 10, issue 3, p. 215-220. Available from: http://www.sciencedirect.com/science/article/pii/S1353829203000674 [Accessed 2015-02-01].

Author **Supervisor**

1 Introduction

Allergology and clinical immunology (ACI) is a branch of Czech health care that aims to prevent and treat diseases induced by immune mechanism disorders or by pathological conditions influenced by immune mechanisms. The most common of these diseases are by far allergy itself and asthma which is often induced by allergy as well. Although 30-40% of the Czech population suffers from diseases treated in ACI, no study of geographic variation in Czech ACI has yet been conducted. (Česká společnost alergologie a klinické imunologie, n.d.; ÚZIS ČR, 2013)

The importance of geographic variation studies is discussed among others by OECD (2014) who show that by exploring this variation, it can be found out whether health care is determined by factors related to needs and preferences of patients or by factors related to physicians. In this way, it is possible to uncover the presence of inefficiency or even inequity in health care.

Based on the 2012 district-level data on care utilization and provision, this work aims to study Czech ACI in the framework of geographic variation using methods of spatial econometrics. The first goal of the analysis is to merely compare the utilization and provision of care in Czech ACI as this has not been done before. Therefore the data on utilization and provision of health services are explored in great detail. Moreover, utilization and provision together with the number of physicians are studied in spatial statistics framework, based on which further differences in geographic structure can be found.

The second goal of this work is to attempt to construct two econometric models one of which would explain care utilization, the other one care provision. The purpose of these models is to show what part of the variation in utilization and provision of care can be explained and to identify districts where care utilization or provision is unreasonably high or low. A variety of potential explanatory variables is used to find the best fitting models. The estimation is first carried out using the method of ordinary least squares, which serves as a benchmark against which spatial models are tested. Multiple spatial model specifications are tested for in order to find the best model, including the spatial autoregressive model, the spatial Durbin model, and the spatial error model. As spatial models in general work with interdependence of neighboring observations, if one of these models is significant, policy implications regarding care coordination could be drawn from it.

The present thesis contributes greatly to the current empirical literature as no works on geographic variation in the branch of ACI as a whole have been published before. Moreover, the thesis puts a yet unused ACI dataset into use. Methodologically, the use of spatial econometrics is rare in studies of geographic variation and thus the employment of spatial methods brings a new insight into this issue. Policy implications can be drawn based on estimated models in terms of policy coordination and available data sources.

The final thesis topic differs slightly from the originally proposed topic. This difference is caused by the fact that the data describing care utilization and the number of ACI physicians were not available to the author at the time of writing the thesis proposal.

The present thesis is structured as follows: Section 2 describes existing literature on ACI and geographic variation and then familiarizes the reader with the Czech fee-for-service system and health services in ACI. Section 3 provides a detailed description of the data and their sources used in this thesis. Section 4 follows by a deeper insight into the spatial characteristics of care utilization and provision, which enables us to uncover their geographic properties. Section 5 explains the methodology used in econometric modeling. Section 6 follows by estimation of models for care provision and utilization, interpretation of results of these models and discussing their real-life implications. Section 7 then concludes the most important findings of the thesis.

2 Literature Review

The branch of ACI together with its most common disease of allergy is described in Section 2.1 in order to outline possible determinants and impacts of ACI diseases and allergy in particular. This is, in Section 2.2, followed by a research of existing literature on geographic variation in health care with focus on application of spatial econometrics and on ACI studies. Geographic variation in health care can be related to preferences and needs of patients. However, it is often the case that it is related also to differences in practice styles and/or unequal access to health services (OECD, 2014). The second type of variation is very important to study as it can point to inequity or inefficiency of the whole health care system (OECD, 2014). The use of spatial econometrics helps to incorporate the geographical structure of the data and relationships that might stem from it as well as approximate unobservable latent factors determining health care.

For a better understanding of the functioning of ACI in Czech environment, Section 2.3 provides information on the mechanisms by which the care is financed and provided. Section 2.4 follows by providing information on the specific health services in Czech ACI.

2.1 Literature on Allergology and Clinical Immunology

According to the Czech Society of Allergology and Clinical Immunology, ACI is an interdisciplinary specialty and is both a clinical and laboratory discipline. Its focus is the diagnostics, treatment, and prevention for patients with diseases induced by immune mechanism disorders or by pathological conditions influenced by immune mechanisms (conditions in which immunomodulation is an important part of therapy and prevention). (Česká společnost alergologie a klinické imunologie, n.d.)

The most common diseases treated in ACI in the Czech Republic (CR) are several types of allergic diseases and asthma (including allergy-induced asthma). The rest of diseases is related to a malfunction of immunity or autoimmunity both congenital and acquired. These diseases, despite their lifelong duration, account only for 14.7% of patients in the branch of ACI. Therefore, only asthma and allergy will be focused on in greater detail in this section. (ÚZIS ČR, 2013; Česká společnost alergologie a klinické imunologie, n.d.)

Allergy is an environment-induced disease defined as an immunologically mediated and allergen-specific hypersensitivity of an individual, where hypersensitivity means an abnormally strong response to a stimulus. It is important to realize that a single common determinant of allergy does not exist. Moreover, it is the individual's innate immune system that forms the reaction to potentially allergenic proteins. There are several basic classes of allergens: house dust mite allergens, pet allergens, tree, weed and grass pollen allergens, food allergens, and insect venom allergens. While some of the allergens are associated with only mild symptoms, symptoms of others are severe. Pet allergens are the most common cause of asthma and allergy to several pets can lead to severe problematic asthma. Tree, weed, and grass pollen are allergens that are widely distributed worldwide and are the main cause of seasonal inhalant allergies. Sensitivity to food allergens is significantly correlated with sensitivity to birch-pollen allergens in Europe. (eds. Akdis & Agache, 2014)

Bartra et al. (2007) identify the use of motor vehicles, especially those using diesel fuels, as an additional factor contributing to allergic diseases. Apart from the direct effect of air pollution, there is also an indirect effect of such emissions. When the diesel exhaust particles get into contact with pollen, they facilitate the penetration of the pollen into human airways. Thus allergy can be expected to have high prevalence where there are a lot of plants with high pollen output and a lot of diesel vehicles at the same time.

Allergic diseases have a negative impact on over one billion people worldwide. Moreover, this number is expected to be four billion in 2050s (eds. Akdis & Agache, 2014). According to de Monchy et al. (2013), the number of people who suffer from allergy is continuously increasing also in Europe, now covering about 50% of European population. Allergy is actually one of the most common afflictions in the Western World. Therefore, it is clear that allergic diseases are and will be, on top of their effects on individuals, a large burden to the society. The costs of allergic diseases are both direct and indirect. Direct costs are those connected to the treatment of allergic diseases whereas indirect costs are those resulting from loss of productivity of a large part of the labor force and therefore of the whole economy. Despite the profound effects of allergic diseases on our everyday life, the knowledge of disease mechanisms, prevention, patient care, and social determinants is very limited. (eds. Akdis & Agache, 2014)

The importance of Czech ACI is obvious from ÚZIS ČR (2013) according to which there were 2,270,871 examinations/procedures in ACI in 2012 in the CR.

These were conducted by over 400 physicians together with over 500 non-physician health care professionals. In the same year 888,748 patients were treated which counts for 8.5% of Czech 2012 population (Czech Statistical Office, 2015a). According to Alexa et al. (2015) the CR has in general a higher number of physicians per population than is the European Union (EU) average. This can be one of the reasons why the CR also has lower mortality due to asthma compared to the EU (Alexa et al., 2015).

Warner et al. (2006) stress the importance of ACI as they point out allergic diseases to be the most common cause of chronic illness in developed countries and recommend the treatment of allergy to be a priority. The situation in the CR is no exception as Česká společnost alergologie a klinické imunologie (n.d.) estimates that 30-40% of Czech population suffer from diseases treated in ACI and that this share is continually increasing and expected to continue increasing in the future.

2.2 Literature on Geographic Variation

In general, literature concerning geographic variation in health care is quite extensive in terms of health care utilization and provision. A number of studies applying spatial methods was conducted to explore geographic variation in certain branches of health care. However, literature on geographic variation in ACI is mostly limited to prevalence of diseases only or to a relation of environmental determinants and disease prevalence.

A major geographic variation study, OECD (2014), studies geographic variation in selected health services in 13 countries including the CR and finds out that geographic variation is present both across and within countries. It is found that patients' needs and preferences account only for a part of the variation and that some of the variation is likely to reflect differing practice styles of physicians. This finding is in general supported by Westert et al. (2003) who also find regional differences in health care supply using data on hospital discharge rates in the Netherlands.

State-level variation in health care utilization in Germany is explored by Eibich & Ziebarth (2014a) based on microdata including both individual- and state-level factors. It is found that the main determinants of care utilization are not only individual health status and behavior but also supply factors at state level. Similar evidence is provided by Skinner (2011) who studies geographic variation in health care utilization and spending and also concludes that health care utilization is at least partially supply-driven.

Although it is not very common, there are previous works studying geographic variation in several branches of health care that employ spatial econometrics or spatial statistics methods. Filippini et al. (2009) use a spatial lag model to study local variations in the use of antibiotics in 240 small areas in Switzerland. It is found that the use of antibiotics is heterogeneous and the spatial lag model contributes its variation mainly to socio-economic characteristics of the population, the incidence of infections, price of antibiotics, and local supply of health care. A model using spatial spillovers is employed by Lauridsen et al. (2008) in order to study public per capita pharmaceutical expenditures in 50 provinces in Spain. Even after controlling for the effect of spatial spillovers, heterogeneity still remains among the provinces.

Eibich & Ziebarth (2014b) estimate spatial health effects in Germany using a hierarchical Bayes model. They find clustering as well as strong spatial dependencies on county level. Bruni & Mammi (2015) also discover clustering in their study of variations in hospital care in Spanish districts. In their work, they employ the spatial autoregressive model and the spatial Durbin model to incorporate geographic proximity of institutions.

Nevertheless, none of the previously mentioned works focus on ACI in particular. Existing studies of geographic variation in ACI are focused mainly on prevalence of ACI diseases and their determinants, not on care utilization or provision in general.

The ACI studies closest to the topic of this work, although not covering the whole branch of ACI, follow. Bousquet et al. (2007) study geographic variation of sensitization to most common allergens and show that variation indeed occurs in this area. However, the majority of ACI studies focus on asthma. Tsai et al. (2009) explore the quality of asthma care in emergency departments in the United States of America and discover variations in quality as well as geographic differences. They also find that following guidelines for treatment recommendation might decrease the amount of hospitalizations. Gupta et al. (2008) explore geographic variability of childhood asthma prevalence in Chicago using geocoded data from 105 schools linked into neighborhoods. They find out that the prevalence varies widely while adjacent neighborhoods often show significantly different levels.

As the present-day literature on geographic variation in ACI care in general is very limited, the author hopes to provide an opening study on this topic, moreover, applying spatial econometrics methods that are not commonly used in ACI literature.

2.3 Fee-for-service Payment System in Czech Specialist Ambulatory Care

The understanding of the functioning of the relevant part of the Czech reimbursement system is necessary for further comprehension of the data sources used and also of the logic behind using the data in the model. In particular, this section focuses on the fee-for-service payment method used in Czech specialist ambulatory care which ACI is a part of.

Statutory Health Insurance (SHI) system is used in the CR. Participation in the system is compulsory for virtually the entire population of the CR because of compulsory membership in a health insurance fund which can, however, be chosen freely. The funds are obliged to accept anybody. The health insurance funds collect contributions from employers and employees, self-employed, and individuals without taxable income who are not insured by state. The state pays contributions for the state-insured (e.g. students, children, those on parental leave, those retired, those unemployed, those below the poverty line, prisoners, and asylum seekers). The coverage is thus virtually universal. (Alexa et al., 2015)

In ambulatory care, individual private practice is by far the most common form of work of physicians in the CR. Health care providers are registered by regional authorities and contracted by health insurance funds (a lengthy process). The health insurance funds are in fact the major purchasers of services in health care in the CR based on individual long-term contracts with providers. Non-hospital ambulatory care specialists (on whom this thesis focuses) receive payments based on a capped fee-for-service scheme. (Alexa et al., 2015)

The List of Health Services (LHS) states what health services are to be reimbursed. Although this is a positive list, services that are not on it may still be reimbursed under special circumstances. Ambulatory specialists provide care, which is reimbursed based on the LHS up to a given threshold (above the threshold, lower prices are used). This type of cap was introduced based on past experience: its former absence led to overproduction and a strict cap led to excessive rationing. (Alexa et al., 2015)

In practice, a provider may decide not to be contracted by any health insurance fund and to rather obtain direct payments from patients. This is, however, extremely rare in branches like ACI and such cases are therefore negligible for the purposes of this study. Typically the vast majority of specialists has contracts with all health insurance funds relevant in their region. (Alexa et al., 2015)

The LHS is defined and modified by Ministry of Health of the Czech Republic (MH) decrees (Czech Republic, 2011a for 2012, the year for which data were provided). In the LHS, there is information on each service – most importantly its name and code, number of points obtained for providing the service, and the time the service provision is supposed to take. The number of points obtained for the service reflects the costs and intensity of the service and is therefore the base for reimbursement. The number of points is then multiplied by the value of point also given by MH (Czech Republic, 2011b for the year 2012). This is how the amount reimbursed for a given health service is computed.

In the CR, there is no gate-keeping which means that patients can obtain care directly from a specialist without the need of a referral. In practice, patients do indeed use this possibility. This direct and unlimited access to ambulatory care is highly valued by patients in the CR. (Alexa et al., 2015)

2.4 Health Services in Czech Allergology and Clinical Immunology

Based on the LHS, the provision and utilization of care in ACI is divided between nine services which will be examined in the present thesis. The services provided and utilized in Czech ACI are (Czech Republic, 2011a):

(i) Complex examination by allergist and clinical immunologist,

(ii) Purposeful examination by allergist and clinical immunologist,

(iii) Checking examination by allergist and clinical immunologist,

(iv) Quantitative assessment of nitric oxide in exhaled air,

(v) Specific immunotherapy by allergen,

(vi) Intradermal skin test by allergen,

(vii) Other intradermal test,

(viii) Skin test by allergen – prick test,

(ix) Double-blind placebo-controlled food oral exposition test.

As this is a thesis in economics, the medical details of these services will not be explained here as they are not of importance. On the other hand, economically

important details on these services are provided in Table 2.1. Each service in the LHS is identified by a 5-digit code, which is unique for each service. The points obtained for the service are a base for reimbursement for that service as explained in more detail in Section 2.3. As visible in Table 2.1, the points for the services differ greatly depending on the costs and complexity of a particular service. The LHS also provides a time in minutes that should be standard when providing the service.

Table 2.1: Health services in ACI

Name of the service	Code	Points	Time (min)
Complex examination by allergist and clinical immunologist	27021	474	60
Purposeful examination by allergist and clinical immunologist	27022	242	30
Checking examination by allergist and clinical immunologist	27023	121	15
Quantitative assessment of nitric oxide in exhaled air	27101	405	10
Specific immunotherapy by allergen	27205	78	10
Intradermal skin test by allergen	27210	26	3
Other intradermal test	27220	11	5
Skin test by allergen – prick test	27240	17	2
Double-blind placebo-controlled food oral exposition test	27260	233	60

Source: Czech Republic (2011a).

3 Data Description

This section describes the sources and form of the data used for the analysis together with the basic modifications of the data performed in order to make them useful for the thesis. Firstly, Czech districts and their neighboring relationships are described which enables one to consider and model spatial relationships in the data. Secondly, numbers of allergists and clinical immunologists in those districts are depicted as ACI physicians are the most important providers of care in ACI. Then the crucial data sets on care provision and utilization are explained to the reader so that measures of care utilization and provision can later be used as dependent variables in econometric models. Finally, data on demographic, socio-economic, and environmental characteristics of districts are described as these might influence care utilization and provision and thus need to be controlled for in the models if significant. Section 3.7 then consists of an overview of all used variables.

3.1 Czech Districts and Neighboring Relationships among Them

As all data used in this work are on district level, it feels natural to start the data description by providing information on Czech districts. There are 77 districts in the CR and for statistical purposes they are coded in 6-character codes according to the LAU (Local Administrative Units) standard. The thesis therefore uses the up-to-date LAU standard for district coding. The list of districts together with these codes can be viewed in Table A.1 in Appendix A. For readers not familiar with Czech geography, a map of Czech regions and districts from 2012 is provided in Figure A.1 in Appendix A.

In the CR, patients are eligible to utilize care in any district, not only the district where they are residents. Therefore, it is likely that there are some relations among neighboring districts in terms of care utilization and provision (these will be discussed in detail in further sections).

In order to be able to capture any such relationship, one needs to construct a matrix containing information about contiguity[1] of districts. For the purposes of this

[1] A district is contiguous to another district if the two districts share a border. (LeSage and Pace, 2009)

thesis, the matrix was constructed using the map of Czech districts provided by Czech Statistical Office (2012). In case of uncertainty about contiguity of some districts, detailed information on districts from Czech Statistical Office (2012) was used.

The contiguity matrix W^* is a 77×77 matrix of zeros and ones. The rows and columns represent districts ordered by their code (which is the same ordering as for all other variables and can be viewed in Table A.1 in Appendix A). The element w_{ij}^* with $i, j = 1, \dots, 77$ of matrix W^* is equal to one if and only if district i is contiguous to district j. If the two districts are not contiguous, its value is zero. Please note two important properties of matrix W^*. First, the matrix is symmetric because whenever district i is contiguous to district j it also holds that district j is contiguous to district i and therefore $w_{ij}^* = w_{ji}^*$ for all $i, j = 1, \dots, 77$. Second, no district is contiguous to itself which means that for all $i = 1, \dots, 77$ the value of w_{ii}^* is zero. In other words, matrix W^* has zeros on the diagonal. This is a common convention in spatial econometrics (LeSage & Pace, 2009).

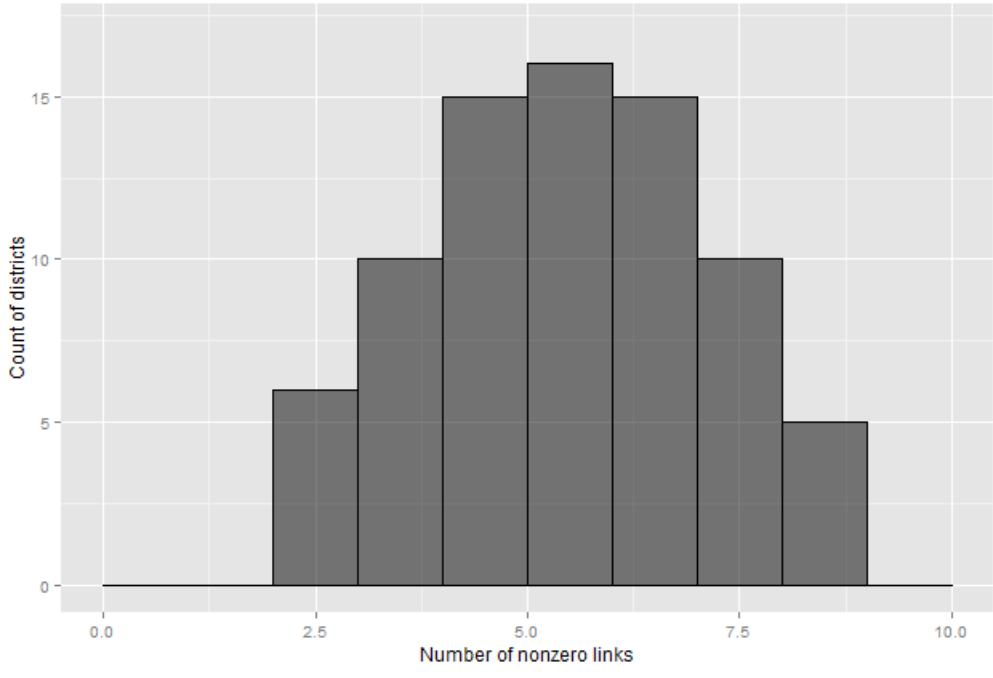

Figure 3.1: Histogram of nonzero links in W*

Source: Author's computations.

The exploration of the links among districts captured by matrix W^* follows. The total number of nonzero links[2] in the matrix is 382 (counting both sides of the

[2] Number of the elements of matrix W^* that are not zero.

symmetric relationship) which accounts for 6.44% of the total of 5,929 elements of the matrix. The number of nonzero links for each district can be computed as a sum of all elements in the corresponding row of matrix W^*. The average number of nonzero links is 4.96 while the median number of nonzero links is 5 (in 16 districts). The lowest number of nonzero links in the dataset is 2 (for districts Prague, Český Krumlov, Sokolov, Jablonec nad Nisou, Brno-town, and Jeseník). On the other hand, the maximum number of nonzero links is 8 (for districts Prague-east, Příbram, Plzeň-south, Plzeň-north, and Louny). For a more clear idea about the contiguity structure, a histogram of the number of nonzero links can be viewed in Figure 3.1.

Thus a binary contiguity matrix W^* was constructed and used for assessment of neighboring relationships among districts. A contiguity matrix is a fundamental part of each spatial econometrics model and test. However, for the purposes of spatial econometrics, a row-standardized version of matrix W^* needs to be created. The row-standardized contiguity matrix W was created from matrix W^* by dividing each element of matrix W^* by the sum of all elements in the corresponding row. Therefore the sum of each row of matrix W is one. Matrix W is not symmetric but it is similar to symmetric[3]. The neighboring relationships are unchanged by this transformation. The row-standardization is a useful transformation because many tests in spatial econometrics work with row-standardized matrices (Bivand et al., 2008).

3.2 Number of Allergists and Clinical Immunologists

The number of physicians is an important indicator of the supply side of the market for health care. The dataset contains average annual number of allergists and clinical immunologists in 2012 for each LAU-coded district in the CR.

Part of the data can be found in the archive of information on health care by the Institute of Health Information and Statistics of the Czech Republic (IHIS), separately in the report for respective region. Kašková (2013a, 2013b, 2013c, 2013d) provides the data for districts in regions Královéhradecký, Liberecký, Pardubický, and Ústecký respectively. Data for districts in regions Jihomoravský, Vysočina, and Zlínský can be found in Pazourková (2013a, 2013b, 2013c) respectively. Data for districts in regions Moravskoslezský and Olomoucký are provided by Kozlíková (2013a, 2013b) respectively.

[3] Matrix W is similar to symmetric because it has the same eigenvalues as matrix W^* which is symmetric. (Bivand et al., 2008)

Data for the rest of the regions (Prague, Středočeský, Jihočeský, Plzeňský, and Karlovarský) are not publicly available on the IHIS website. Nevertheless, the author obtained them upon request by e-mail from IHIS employee (Nováková, 2015).

The 2012 average annual number of ACI physicians per 100,000 inhabitants[4] takes values from 0 in districts Prachatice and Plzeň-north to 11.63 in the district Brno-town. The average annual number of physicians per 100,000 inhabitants of district is 3.16, the median value is 2.80. The data are positively skewed (skewness parameter 2.03) which can be seen in Figure 3.2. In the models, the variable is denoted $PhysPC$.

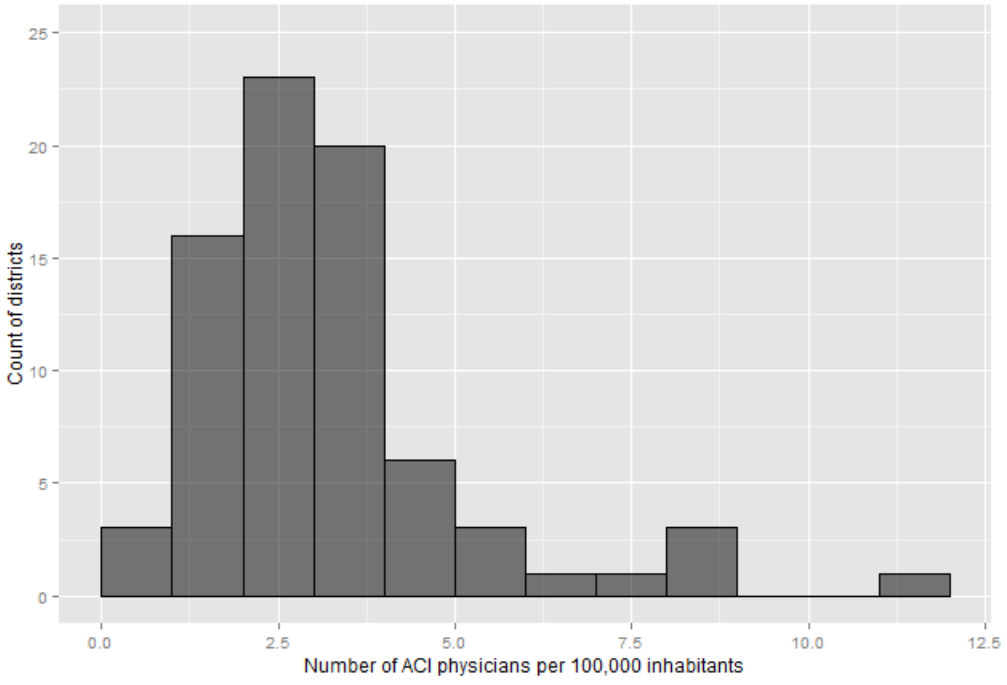

Figure 3.2: Histogram of ACI physicians in districts

Source: Author's computations.

Czech patients can, however, visit not only physicians in the district of their residence but also physicians in other districts. As there is obvious shortage of allergists and clinical immunologists in at least some districts (zero ACI physicians in two districts), it is likely that people in some districts will seek care in neighboring districts. Two measures for the number of physicians in neighboring districts are constructed.

[4] The number of inhabitants in districts was obtained from Czech Statistical Office (2015a).

The first measure is the number of ACI physicians in all neighboring districts per 100,000 inhabitants. To construct this measure for a district, one uses the contiguity matrix W^* described in Section 3.1 and uses data for districts that are contiguous to the district of interest. The value of this new variable $NeighPhysPC$ for district j can be computed as

$$NeighPhysPC_j = \frac{\sum_{i=1}^{n} Phys_i \cdot w_{ij}^*}{\sum_{i=1}^{n} Pop_i \cdot w_{ij}^*}$$

where $Phys_i$ is the number of ACI physicians in district i, Pop_i is the population[5] of district i, w_{ij}^* is an element of the binary contiguity matrix W^*, and $n = 77$ is the number of districts. In other words, as the first step, one computes the total number of ACI physicians in all neighboring districts. In the second step, one computes the total population in all neighboring districts. Then one divides the number of physicians from step 1 by population from step 2 and obtains the number of ACI physicians in all neighboring districts per 100,000 inhabitants. This is done for all 77 districts. In the models, this variable is denoted $NeighPhysPC$ and its basic descriptive statistics can be viewed in Table 3.1.

Table 3.1: Number of ACI physicians in neigh. districts per 100,000 inhabitants

min	1.14
1st quartile	2.35
median (2nd quartile)	3.05
3rd quartile	3.84
max	6.66
mean	3.24

Source: Author's computations.

The second measure is more straightforward as it is based directly on the number of ACI physicians in district per 100,000 inhabitants. The value of this new measure $AvNeighPhysPC$ for district j can be obtained as

$$AvNeighPhysPC_j = \frac{\sum_{i=1}^{n} PhysPC_i \cdot w_{ij}^*}{\sum_{i=1}^{n} w_{ij}^*}$$

where w_{ij}^* is again an element of the binary contiguity matrix W^*, and $n = 77$ is the number of districts. To construct this measure for a district, one thus simply computes

[5] The population in districts was obtained from Czech Statistical Office (2015a).

the average of the number of ACI physicians per 100,000 inhabitants in all districts contiguous to the district of interest. Thus one obtains the average of the number of ACI physicians per 100,000 inhabitants in neighboring districts. Again, this is done for all 77 districts. This variable is denoted $AvNeighPhysPC$ in the models and its basic descriptive statistics can be viewed in Table 3.2.

Table 3.2: Average number of ACI physicians in neigh. districts per 100,000 inh.

min	1.08
1st quartile	2.35
median (2nd quartile)	2.95
3rd quartile	3.63
max	5.52
mean	3.00

Source: Author's computations.

3.3 Amount of Services Provided

The base of the analysis of care provision as well as the dependent variable of the model describing care provision are data on the actual amount of health services provided in ACI in the CR. The dataset from which the data were extracted was provided by the MH according to the intra-institutional agreement nr. 0317/13 between the MH and the Charles University in Prague. Detailed information on the dataset can be found in Troch (n.d.).

The MH database provided to the author contains data on all health services provided to the insured by health care providers in the year 2012 in the CR. Please note that the dataset provided to the author was pre-aggregated by health care provider and health insurance fund and thus information on amounts of services related to particular providers or funds are not included in the dataset. The database is grouped by half-year, district, category of health care provider, specialty of the physician, and health service code. For each of the groups, there is the amount of actual services provided in 2012 and also the number of points related to these services. Please note that the half-year and category of health care provider are not useful for the analysis as data are not informative and are therefore not used any further so the final version of the data used is grouped by district, specialty of the physician, and health service code only.

The standard of coding of statistical areas of the CR has changed several times during the past decade. Despite the fact that this should not be a problem when

using data for only one year, apparently not all health care providers updated their systems and the database, therefore, still contains some outdated codes. Fortunately, this problem is easy to fix using the converter provided by Český statistický úřad (2014), so that the data are compatible with the LAU codes described in Section 3.1.

For the purposes of the analysis conducted in this thesis it is absolutely crucial to realize that in the dataset, the care provision is related to a district if and only if the care was provided by a physician registered in that district regardless of residence of the patient to whom the care was provided. This means that if a patient who is a resident at district A visits a physician in district B, then the care provided in this case will be related to district B (not A) in the dataset for care provision. The importance of this convention will be evident when its counterpart for utilization of care is described in Section 3.4.

The specialty of the physician contains information on what type of physician provided the health service. Here it is crucial to realize that not only allergists and clinical immunologists may provide care classified as ACI. Actually, many other types of physicians provide these services on a daily basis. In 2012, 56 types of physicians including allergists and clinical immunologists provided care in ACI. The list of all 324 specialties together with their Czech names can be downloaded from Všeobecná zdravotní pojišťovna České republiky (2010). In order to maintain homogeneity of the dataset, services provided as outpatient hospital care are not included in the analysis (the filtering is based on their different codes). Still, more than 99.9% of the original dataset is used in terms of the quantity of services provided.

When dealing with ACI, it is necessary to use only the part of the database which is relevant to it. Therefore, only those services which are part of the specialty nr. 207 (ACI) are to be included. The nine ACI services were described in detail in Section 2.4.

Some conclusions can already be drawn from examining basic descriptive statistics of the data only. In Table 3.3, the summary of the amount of services per 100,000 inhabitants[6] provided in districts for each of the nine ACI services can be found. The minimum is zero for all services because no care in ACI was provided in Prachatice district. This can be a result of either a mistake in the data or of the fact that there are no ACI physicians in the district (as already shown in Section 3.2). The

[6] Population of districts was obtained from Czech Statistical Office (2015a).

most common services according to the statistics in Table 3.3 are: skin test by allergen – prick test (several times higher mean and median than other services), purposeful examination by allergist and clinical immunologist, and checking examination by allergist and clinical immunologist. On the other hand, other intradermal test and double-blind placebo-controlled food oral exposition test are the services that are least common, the former being provided in only about a half of the districts, the latter being provided in four districts only.

Table 3.3: Provision by district (per 100,000 inhabitants) - descriptive statistics

Name of the service	Min	Max	Mean	Median
Complex examination by allergist and clinical immunologist	0.00	2,594.36	1,024.92	976.94
Purposeful examination by allergist and clinical immunologist	0.00	14,726.15	4,571.38	4,079.83
Checking examination by allergist and clinical immunologist	0.00	14,147.57	5,624.32	4,840.54
Quantitative assessment of nitric oxide in exhaled air	0.00	7,205.58	1,425.68	1,076.18
Specific immunotherapy by allergen	0.00	904.00	244.04	185.28
Intradermal skin test by allergen	0.00	27,325.48	1,627.63	15.13
Other intradermal test	0.00	4,103.04	452.10	0.89
Skin test by allergen - prick test	0.00	54,166.51	22,287.94	19,393.69
Double-blind placebo-controlled food oral exposition test	0.00	4.92	0.12	0.00

Source: Author's computations.

The following analysis, however, considers the branch of ACI as a whole and therefore uses the sum of number of points obtained for provision of all nine services in each district. This is a good way of defining the overall care provision as the number of points indicates both the complexity and costs of the service. Therefore, the points are a better measure than the amount of services provided when one combines the nine services into a single measure. Czech Statistical Office (2015a) is

used for transforming the variable into per capita terms. The number of points per capita for services provided in ACI is denoted $ProvPC$ in the models.

Descriptive statistics of this new variable can be viewed in Table 3.4. The minimal value of points for services provided is again 0.00 as no care was reported to have been provided in Prachatice district.

Table 3.4: Overall provision of ACI care in districts per capita

min	0.00
1st quartile	30.01
median (2nd quartile)	40.16
3rd quartile	52.78
max	122.45
mean	42.93

Source: Author's computations.

To get an overview of the care provided in the branch of ACI the histogram of points per capita in districts is provided in Figure 3.3. The data exhibit a positive skew of 0.96 with an obvious outlier represented by the district Brno-town with 122.45 points per capita.

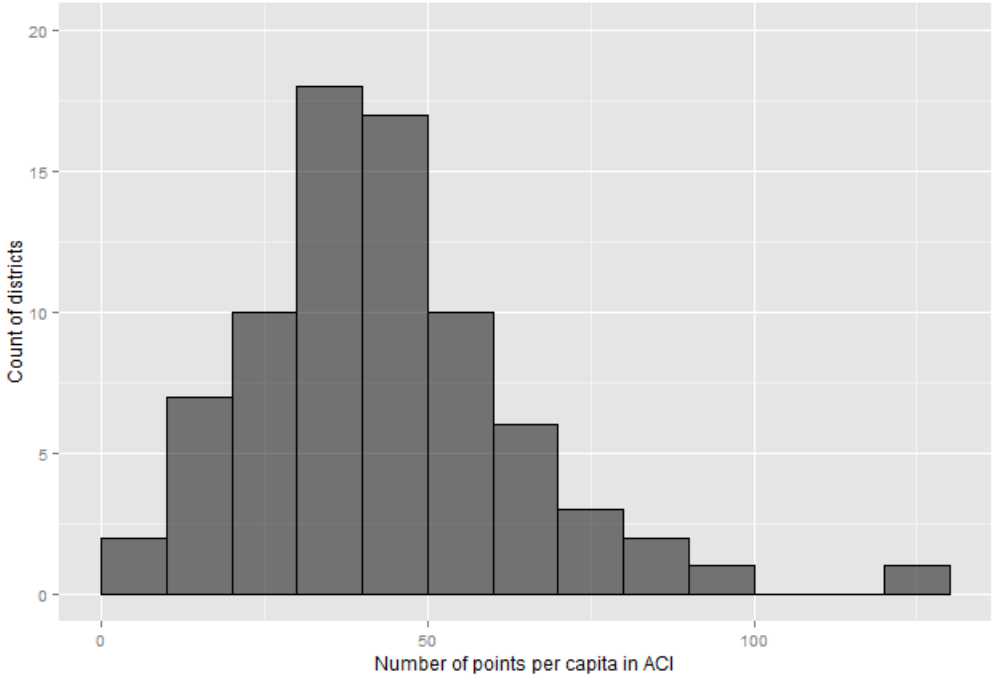

Figure 3.3: Histogram of points provided in ACI per capita in districts

Source: Author's computations.

In essence, however, it is obvious that, because the care provision is bound to the physician providing the service in terms of the district to which the service is then counted, this relationship is more or less tautological. In other words, where there are physicians, there is service provision and vice versa. Statistically measured, the correlation of points for care provision and the number of ACI physicians is 0.99.

Table 3.5: Points for ACI services provided per ACI physician

min	646,654
1st quartile	1,159,993
median (2nd quartile)	1,317,558
3rd quartile	1,662,235
max	5,094,868
mean	1,474,819

Source: Author's computations.

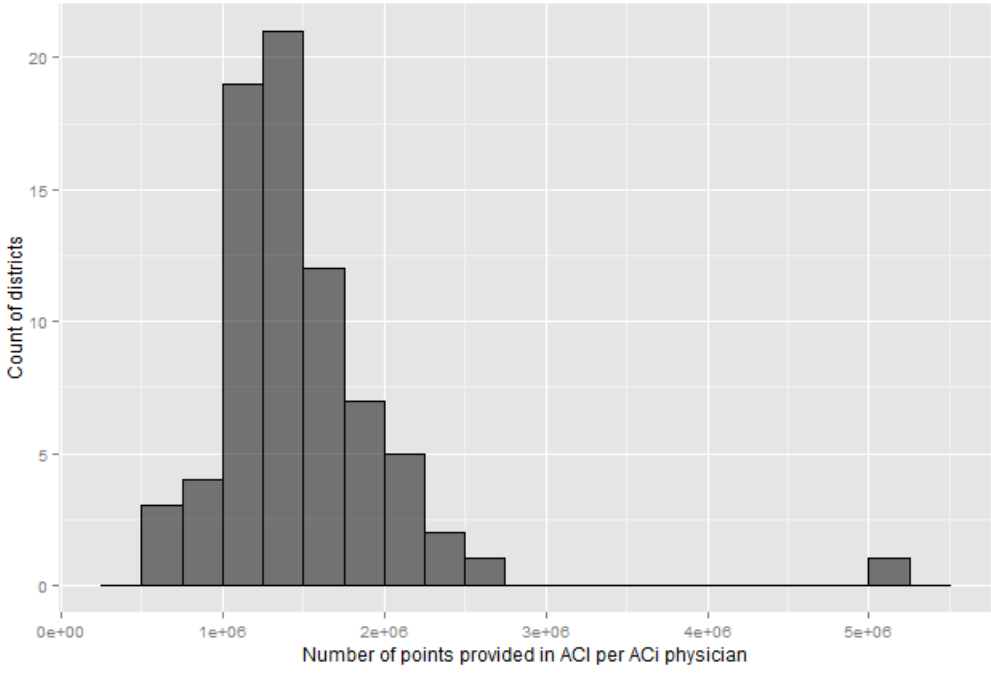

Figure 3.4: Histogram of points provided in ACI per ACI physician

Source: Author's computations.

Therefore, it makes sense to create a variable for care provision which when used, would allow for proper modeling. One such way is to model not points per capita but points per ACI physician instead. $ProvPP$, the number of points for ACI services provided per ACI physician was obtained simply by dividing the overall number of points for ACI services by the number of ACI physicians in each district.

As there is zero number of physicians in two districts, $ProvPP$ is not defined in these two districts (Prachatice and Plzeň-north) and it therefore contains only 75 observations. Basic descriptive statistics of care provision per ACI physician are shown in Table 3.5. The variable also shows a positive skew, the skewness statistic is 3.23. A histogram of this variable can be seen in Figure 3.4, where the skew is visible. The obvious outlier with 5,094,868 points per physician is the district of Třebíč, which in fact did not show any extreme values in care provision or number of physicians itself.

One of the variables that could be related to how many services (in terms of points) physicians provide, is the fraction of points in that district that were actually provided by ACI physicians. If patients visit e.g. a general practitioner instead, the ACI physician would, ceteris paribus, provide less care. The number of points for services provided in ACI by ACI physicians as a percentage of all points for services provided in ACI is denoted $pctgACI$ and was computed using codes for specialty of the physician from the dataset on care provision. The variable is not defined for Prachatice district as no care was reported to have been provided there. The mean value for the percentage of services provided by ACI physicians is 86.03%, the minimum is 18.36% in the district Mladá Boleslav, and the maximum is 100% in Plzeň-town. The histogram of this variable can be seen in Figure 3.5.

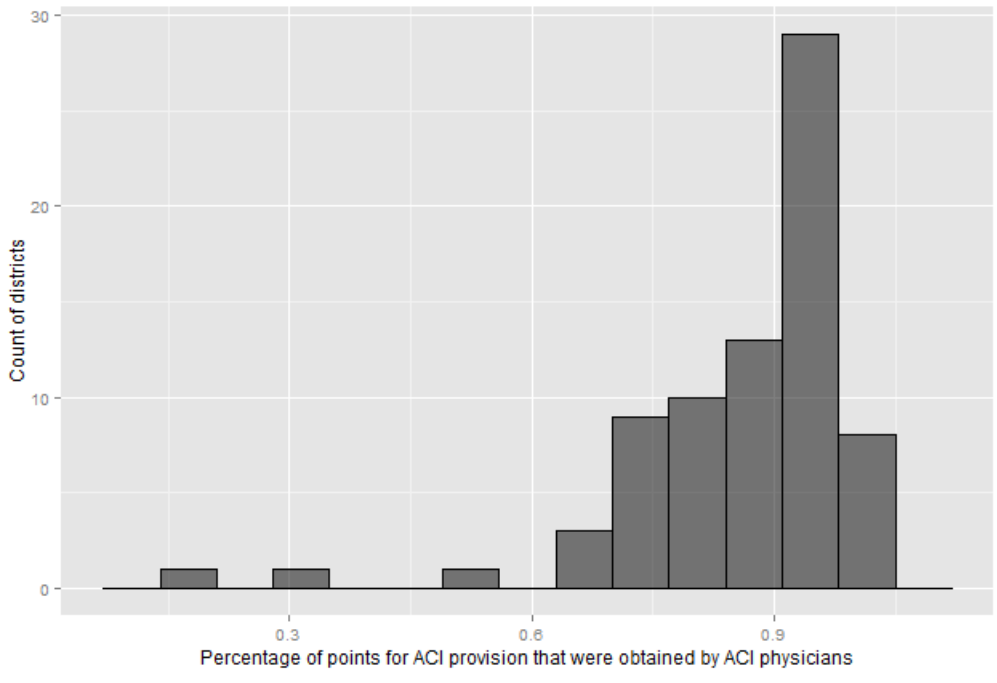

Figure 3.5: Percentage of points in ACI provided by ACI physicians

Source: Author's computations.

3.4 Amount of Services Utilized

The amount of services utilized in Czech ACI in 2012 is crucial for the analysis of utilization of care as it serves as the dependent variable of the model. The reader will find that in many ways the dataset of utilization of care is an analogy to the one for care provision. The dataset for utilization of care was also obtained from MH based on the intra-institutional agreement mentioned in Section 3.3 and its detailed description can be found in Troch (n.d.) as well. Please note that the dataset provided to the author again does not contain information on amounts of services related to particular health care providers or insurance funds.

The database for utilization of care contains aggregated data on all utilization of health services by the insured in the CR during the year 2012. The data in this database are grouped by district, half-year, category of health care provider, and health service code. For each of these groups the database contains a number of services that were utilized. As in the case of care provision, the category of health care provider and the half-year are dropped from the analysis as they are of no real use for the purposes of this study. The resulting dataset therefore includes district, health service code, and number of services utilized only. Please note that the coding of districts is identical to the LAU coding described in Section 3.1.

The fundamental distinction between the two datasets is that (unlike in the case of care provision) in the case of utilization of care the district, which the care utilized is related to, is determined by the residence of the patient. Hence if a patient who is a resident at district A visits a physician in district B, the care utilized in this case will be related to district A (not B, which would be the case for care provision as explained in Section 3.3). Therefore, a difference between the two datasets will occur whenever a patient utilizes care in a district different from the one at which they are a resident.

The services utilized in ACI are identical to those provided. Similarly as in the case of care provision, only the part of the dataset containing data on services in ACI is used. The list of these nine ACI services together with their codes, points obtained for them, and estimated time they are supposed to take was provided in Table 2.1 in Section 2.4.

Table 3.6 shows basic descriptive statistics of the number of services utilized per 100,000 inhabitants[7] by district for each of the nine ACI health services. Unlike

[7] Population of districts was obtained from Czech Statistical Office (2015a).

in the case of provision of care, at least some care was utilized in all districts and the minima are positive except for the double-blind placebo-controlled food oral exposition test which was utilized by patients from 11 districts only (this also causes its zero median). Other intradermal test, although it is not very commonly used, was utilized in all districts (unlike in the case of provision where it was related only to some districts). The most frequently utilized service is by far the skin test by allergen - prick test. Purposeful and checking examinations by allergist and clinical immunologist are also very commonly utilized.

Table 3.6: Utilization by district (per 100,000 inhabitants) - descriptive statistics

Name of the service			
Min	Max	Mean	Median
Complex examination by allergist and clinical immunologist			
379.46	1,843.74	1,183.21	1,152.38
Purposeful examination by allergist and clinical immunologist			
1,853.28	11,084.08	5,345.35	5,159.91
Checking examination by allergist and clinical immunologist			
1,940.83	14,301.02	6,228.48	5,925.68
Quantitative assessment of nitric oxide in exhaled air			
121.08	6,935.27	1,609.39	1,339.22
Specific immunotherapy by allergen			
54.02	836.52	263.97	207.85
Intradermal skin test by allergen			
59.95	25,710.27	1,941.40	491.28
Other intradermal test			
11.87	3,020.73	513.81	140.27
Skin test by allergen - prick test			
7,539.54	51,583.22	25,220.09	24,330.21
Double-blind placebo-controlled food oral exposition test			
0.00	3.21	0.25	0.00

Source: Author's computations.

When one multiplies the count of services utilized by the number of points corresponding to those services and sums the nine quantities up for each district, one can again construct a measure of overall utilization of care in Czech ACI in 2012. In order to transform this measure into per capita terms, population of districts from Czech Statistical Office (2015a) is used. The number of points for services utilized per capita is denoted $UtilPC$ in the models. The descriptive statistics of this variable

are provided in Table 3.7. The same measure is graphically represented by the histogram provided in Figure 3.6. Like the corresponding measure for provision, it has a positive skew of 0.39.

Table 3.7: Overall utilization of ACI care in districts per capita

min	18.71
1st quartile	29.69
median (2nd quartile)	37.43
3rd quartile	44.99
max	59.82
mean	37.65

Source: Author's computations.

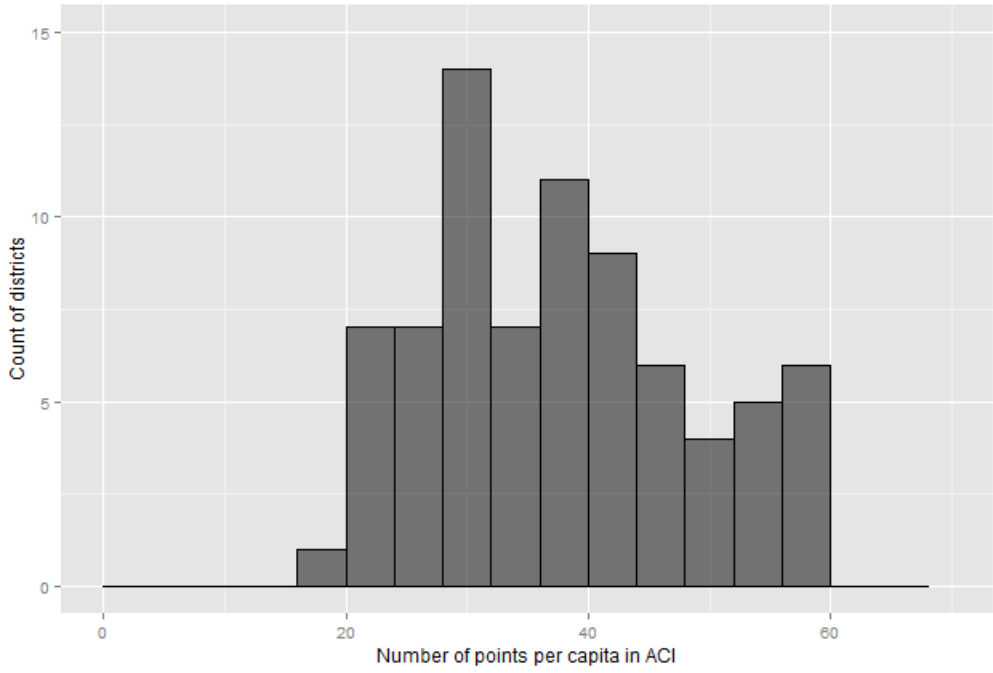

Figure 3.6: Histogram of points utilized in ACI per capita in districts

Source: Author's computations.

3.5 Demographic and Socio-Economic Indicators

Apart from the health care data on ACI, it is necessary for the study to incorporate also other variables which might determine the demand for ACI care. The first group of these variables is demographic and socio-economic indicators.

The population in districts is a variable necessary for conversion of data in absolute units into per capita terms. The data on population in districts in the year

2012 were obtained from Czech Statistical Office (2015a). The number of inhabitants ranges from 40,189 in Jeseník to 1,246,780 in Prague. Therefore it indeed is crucial to use variables in per capita terms for the analysis as the absolute values are not comparable. The mean district population is 136,573.

Although the demand for health care is in general increasing with age as claimed e.g. by European Commission (2015), the branch of ACI is an exception from this rule. In 2012, 49% of patients who received services in ACI were 19 years old or younger (ÚZIS ČR, 2013). Therefore one can expect the age structure in a district to have an impact on the amount of care that patients need. Thus two measures of age structure by Czech Statistical Office (2015b) for the year 2012 are used. First, the index of ageing provides the ratio of people aged 65 years or more to those aged 14 years or less times 100. This index takes values from 69.7 in Prague-west to 137.5 in Plzeň-town, the average value being 111.4. The index of ageing is denoted $OAindex$ in the models. Second, the average age by district (denoted $AvAge$ in models) can also be used to describe the population age structure. It ranges from 38.3 years in Prague-east and Prague-west to 42.5 years in Plzeň-town with the average value of 41.2. As the two indices are highly correlated (correlation coefficient is 0.98) it is clear that they cannot be both used in the same model.

The economic situation is also likely to affect the need for health care in ACI as a worse economic situation usually results in consumption of food of lower quality which is likely to induce more allergic reactions according to ČPZP (2009). For this purpose, the economic situation in districts is proxied by the 2012 average annual unemployment rates obtained from Český statistický úřad (2015) and denoted $Unemp$ for modeling purposes. The lowest unemployment rate is observed in Prague-east and takes value 2.87%, the highest one on the other hand is 12.04% in Most. The average value over all districts is 7.00%.

Population density is also an important demographic indicator as one can distinguish between rural and urban districts using it. Population density for each district was computed using the population in districts (Czech Statistical Office, 2015a) divided by the district area (Czech Statistical Office, 2015c). Population density is denoted $Pdens$ in the models and it ranges from 0.37 (Prachatice) to 25.13 (Prague) inhabitants per hectare. The mean value of $Pdens$ is 1.95 inhabitants per hectare.

3.6 Environmental Characteristics

As already discussed in Section 2.1, environment is a crucial determinant of prevalence of allergic diseases. However, there is no sole environmental determinant of allergy and therefore several environmental measures will be used in the models.

The presence of pollen particles is deemed to be dependent on the fraction of district area where plants grow. On the other hand, water bodies do not produce any pollen at all. Therefore the division of the total district area into several different types of land provided by Czech Statistical Office (2015c) will be used in the model. The district area is divided between arable land, gardens and orchards, permanent grassland, hop gardens and vineyards, forest land, water body areas, built-up areas and yards, and other areas. The descriptive statistics of shares for each of these area types can be found in Table 3.8.

Table 3.8: Environmental characteristics of districts

variable	minimum	maximum	mean	median
District area [ha]	23,019	194,565	102,424	99,271
Arable land [%]	5.6	66.1	37.8	39.9
Gardens and orchards [%]	0.7	9.9	3.1	2.7
Permanent grassland [%]	1.4	29.6	12.0	12.3
Hop gardens and vineyards [%]	0.0	8.9	0.4	0
Forest land [%]	10.3	59.5	32.6	31.4
Water body areas [%]	0.8	6.8	2.0	1.7
Built-up areas and yards [%]	0.5	10.0	2.0	1.6
Other areas [%]	5.3	36.8	10.1	7.9
Emissions				
Solids [t/km^2]	0.1	3.2	0.4	0.3
Sulphur dioxide [t/km^2]	0.0	54.7	3.0	0.5
Nitrogen oxides [t/km^2]	0.1	38.1	2.3	0.4
Carbon monoxide [t/km^2]	1.1	154.4	6.7	3.9
Protected areas [%]	0.1	75.4	16.4	7.7
Roads Length [m/ha]	1.7	11.1	7.4	7.2

Note: % sign indicates percentage of total district area.
Source: Author's computations.

Emissions of pollutants are also likely to have an adverse effect on allergic diseases. Therefore emissions of solids, sulphur dioxide, nitrogen oxides and carbon monoxide (Czech Statistical Office, 2015d) are to be included in the model. The measure used are tons per kilometer squared rather than tons only in order to incorporate the polluting effect and adjust for different district areas. For details about emission values see Table 3.8.

Protected areas as a share of total district area are also to be included in the model as the purpose of protected areas is the maintenance of good environmental conditions. The total area of protected areas for each district was obtained from Czech Statistical Office (2015e) and the recalculation into shares of total area was performed using the district areas provided by Czech Statistical Office (2015c). The share of protected areas differs greatly across districts as can be seen in Table 3.8.

Finally, the length of roads per hectare will be included in the model. This is firstly because it can approximate the amount of diesel pollutants and secondly because it also shows how close one is likely to live to a road. The length of roads in meters per hectare of district area was computed using the length of roads in kilometers (Czech Statistical Office, 2015f) and the total district area (Czech Statistical Office, 2015c). The unit of meters per hectare was chosen due to scaling advantages. As for all other environmental measures, its descriptive statistics can be found in Table 3.8.

3.7 Overview of Used Variables

A number of variables from multiple data sources was introduced to the reader in previous parts of Section 3. In order to ease the reader's orientation in further parts of the text, a summary of notation, descriptions, and values of all used variables is provided in Table 3.9.

Table 3.9: Overview of used variables

Variable notation	Variable description	Values
W^*	binary contiguity matrix	{0,1}
W	row-standardized contiguity matrix	[0.0,0.5]
$PhysPC$	number of ACI physicians per 100,000 inhabitants	[0.00,11.63]
$NeighPhysPC$	number of ACI physicians in all neighboring districts per 100,000 inhabitants	[1.14,6.66]
$AvNeighPhysPC$	average number of ACI physicians per 100,000 inhabitants in neighboring districts	[1.08,5.52]
$ProvPC$	number of points per capita for provision of ACI services	[0.00,122.45]
$ProvPP$	number of points per ACI physician for provision of ACI services	[646654,5094868]
$pctgACI$	percentage of points for ACI provision that were provided by ACI physicians	[0.18,1.00]
$UtilPC$	per capita number of points for utilization of ACI services	[18.71,59.82]

$OAindex$	index of ageing ("over 65 / under 14" ·100)	[69.7,137.5]
$AvAge$	average age	[38.3,42.5]
$Unemp$	average annual unemployment rate	[2.87,12.04]
$Pdens$	population density (inhabitants per hectare)	[0.37,25.13]
$envir_1$	arable land as percentage of district area	[5.6,66.1]
$envir_2$	gardens and orchards as percentage of district area	[0.7,9.9]
$envir_3$	permanent grassland as percentage of district area	[1.4,29.6]
$envir_4$	hop gardens and vineyards as percentage of district area	[0.0,8.9]
$envir_5$	forest land as percentage of district area	[10.3,59.5]
$envir_6$	water body areas as percentage of district area	[0.8,6.8]
$envir_7$	built-up areas and yards as percentage of district area	[0.5,10.0]
$envir_8$	other areas as percentage of district area	[5.3,36.8]
$envir_9$	emissions of solids [t/km^2]	[0.1,3.2]
$envir_{10}$	emissions of sulphur dioxide [t/km^2]	[0.0,54.7]
$envir_{11}$	emissions of nitrogen oxides [t/km^2]	[0.1,38.1]
$envir_{12}$	emissions of carbon monoxide [t/km^2]	[1.1,154.4]
$envir_{13}$	protected areas as percentage of district area	[0.1,75.4]
$envir_{14}$	roads length [m/ha]	[1.7,11.1]

Note: {·} denotes a set, [·,·] denotes a closed interval.
Source: Author.

4 Data in Spatial Context

The purpose of this section is twofold. First, graphical visualization of data on care provision and utilization and on number of physicians will be provided together with their comparisons. Second, this section will analyze spatial autocorrelation in used variables as a foundation for econometric modeling.

The visualization of variables and statistics related to them would not be possible without the spatial and statistical map of Czech districts. The shapefile used for this visualization can be downloaded (upon free registration) from URRlab (2013), which is a project of the Faculty of Science of Charles University in Prague. Despite the fact that the shapefile provided is for the year 2011, it still can be used for visualization purposes even though the author is aware that minor changes in boundaries of districts took place between 2011 and 2012. That is why the shapefile was not used for creation of the contiguity matrix described in Section 3.1 and is used for visualization only.

For the purpose of better understanding of spatial relationships, it is essential to define first the terms *spatial data*, *spatial dependence*, *spatial lag*, and *spatial autocorrelation*. Spatial data samples are samples whose observations are points or regions in space (LeSage & Pace, 2009), in the case of the data used for this thesis, the observations are Czech districts. Spatial dependence often occurs in spatial data samples. Spatial dependence exists when values observed for one location depend on values observed in nearby locations (LeSage & Pace, 2009). In the case of the present thesis, spatial dependence arises when the value observed in one Czech district depends on values observed in neighboring districts. According to Bivand et al. (2008), the spatial lag of any variable y for observation i is $\sum_{j=1}^{n} w_{ij} y_j$ where w_{ij} are elements of a row-standardized contiguity matrix (W in the case of this thesis). In this case, for each district the spatial lag represents a linear combination of values of y constructed from neighboring districts. Spatial autocorrelation refers to correlation of a variable and its spatial lag (Bivand et al., 2008).

The analysis of spatial autocorrelation is based on the Moran's I test introduced by Moran (1948) as it is the most common test for spatial autocorrelation according to Bivand et al. (2008). There are two versions of the Moran's I test: global and local. The global Moran's I test simply tests the spatial autocorrelation of a variable, its null hypothesis is no spatial autocorrelation and the most commonly used

alternative hypothesis is positive spatial autocorrelation. The local version of the test, which basically breaks down the relationship of the variable and its spatial lag into components for observations, is designed to detect clusters and hotspots in data. Clusters are observations with similar neighbors and significantly positive local Moran's I statistic, whereas hotspots are observations that differ from their neighbors and show a significantly negative local Moran's I statistic. The p-value for both tests is derived based on divergence of the statistic from its expected value under normality. The row-standardized contiguity matrix W is used for both tests. (Bivand et al., 2008)

The next concern of this section will be the analysis of care utilization and provision in space. The natural first step is to compare the values of points for care provision and utilization in districts. For the sake of comparison, Figure 4.1 shows the division of Czech districts between those where provision of care is greater than utilization and vice versa. There is no district where these two values would be identical.

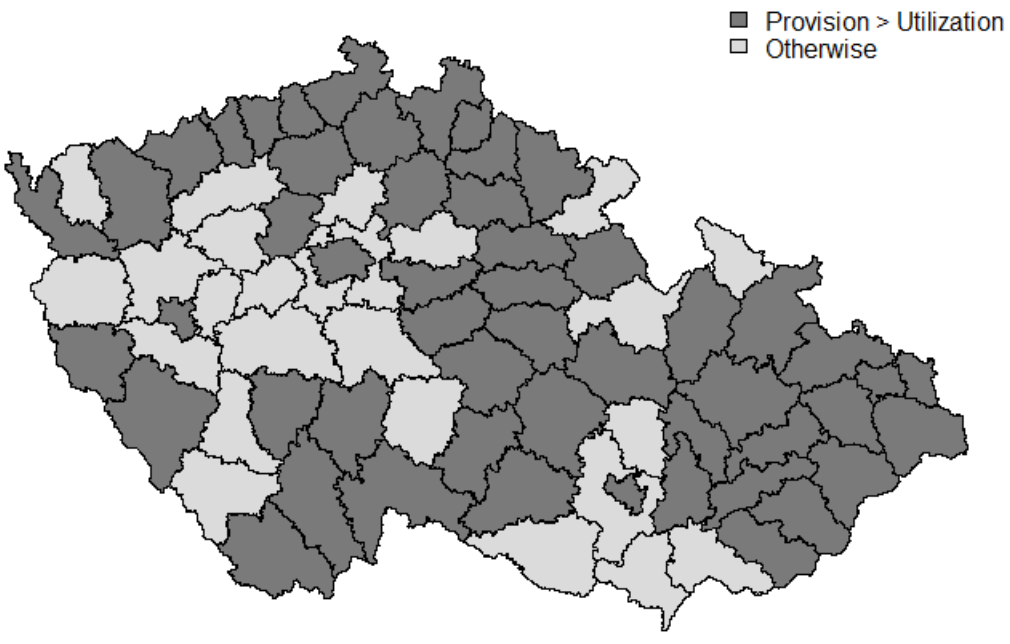

Figure 4.1: Comparison of provision and utilization (map)

Source: Author based on shapefile by URRlab (2013).

Figure 4.2 shows utilization per capita ($UtilPC$) in a map. As seen in the map, the data seem to be spatially correlated, which is confirmed by the global Moran's I test with p-value $3.52 \cdot 10^{-9}$ which indicates a strong rejection of the test's null hypothesis of no spatial autocorrelation. For a more accurate idea of spatial autocorrelation in care utilization, Figure 4.3 shows the sign of local Moran's I in districts. The local

Moran's I is positive in the majority of districts, which is in accord with the result of the global test: the values are spatially correlated, i.e. values in neighboring districts are similar. In order to be able to draw any conclusions about possible clusters and hotspots in the data, it is necessary to employ statistical significance. Hence Figure 4.4 depicts only those values of local Moran's I that are statistically significant at 5%. As can be seen from the figure, there are four clusters among the districts and there is no significant hotspot.

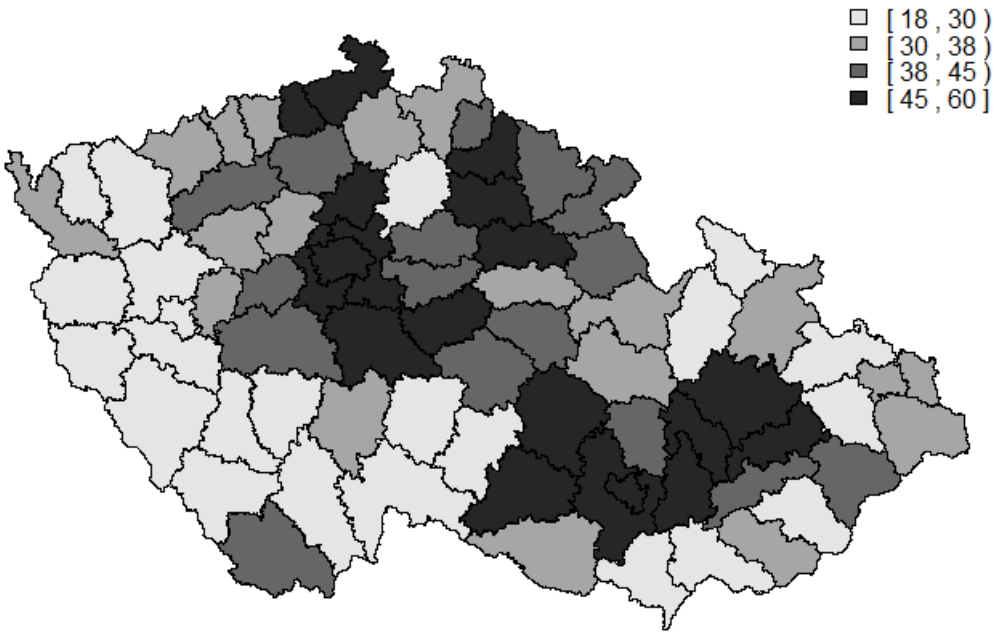

Figure 4.2: Map of points utilized per capita in districts

Source: Author based on shapefile by URRlab (2013).

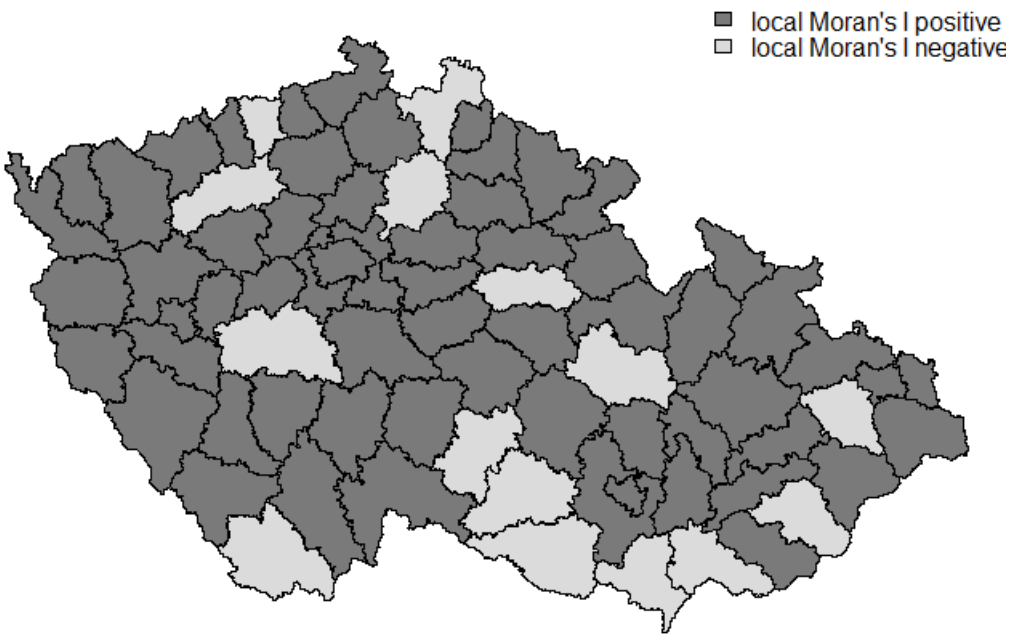

Figure 4.3: Local Moran's I - utilization per capita

Source: Author based on shapefile by URRlab (2013).

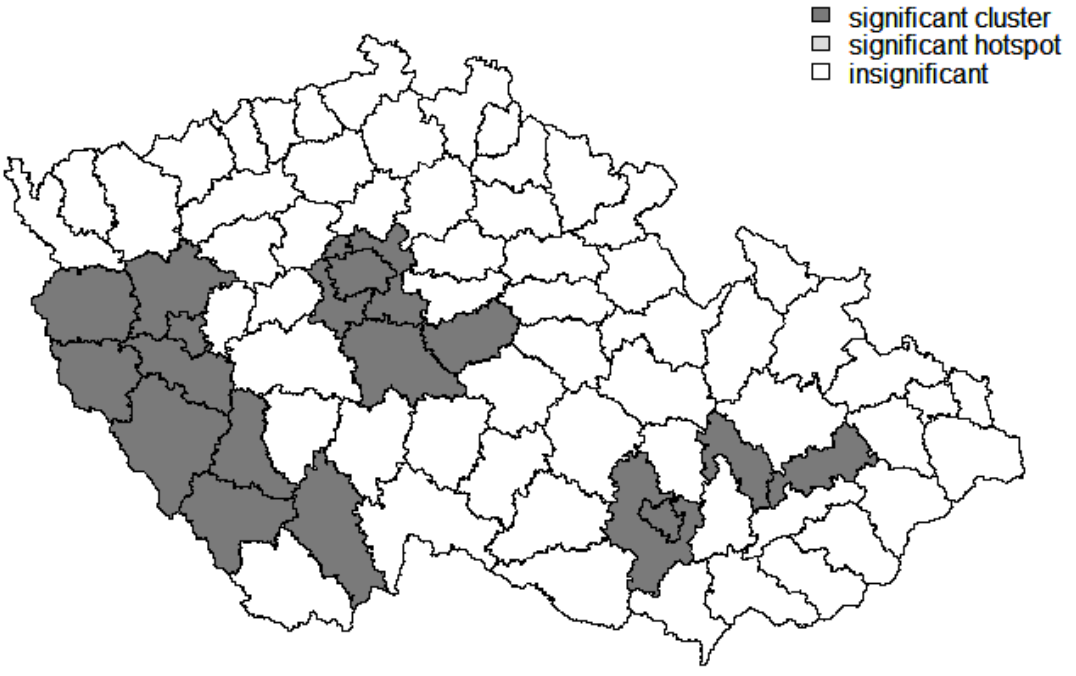

Figure 4.4: Clusters and hotspots in care utilization per capita

Source: Author based on shapefile by URRlab (2013).

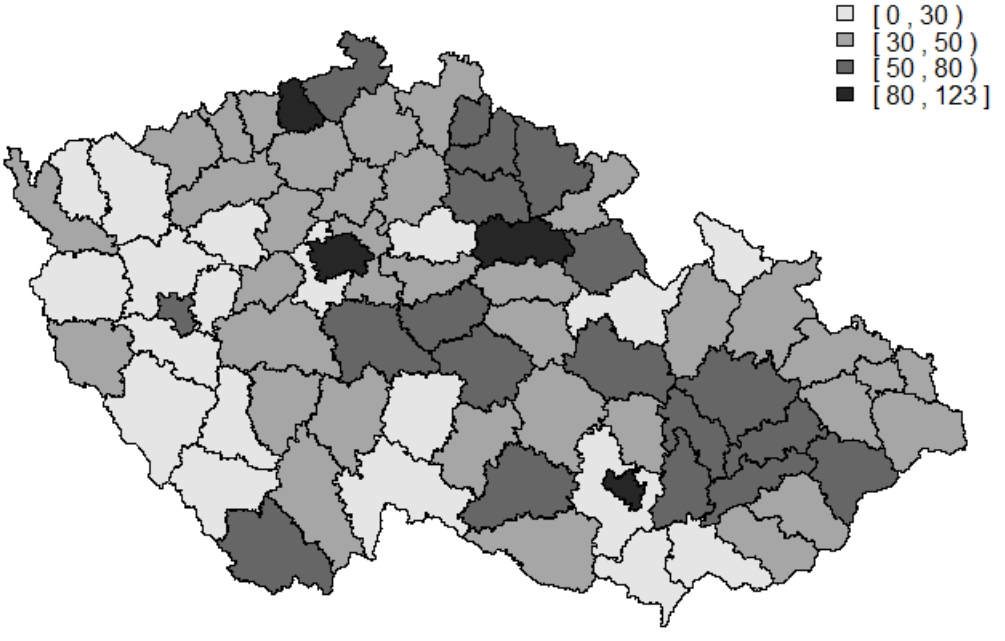

Figure 4.5: Map of points provided per capita in districts

Source: Author based on shapefile by URRlab (2013).

In case of wishing to compare provision and utilization of care in spatial context, one can plot the same three maps for provision per capita ($ProvPC$) as well. Figure 4.5 seems to give a mixed evidence of some seemingly interconnected areas

with a few outlying districts with very high values of provision. The p-value of global Moran's I test for provision per capita is 0.11, therefore one fails to reject the null of no spatial autocorrelation at 5% as well as at 10% level and concludes that the variable is not spatially autocorrelated.

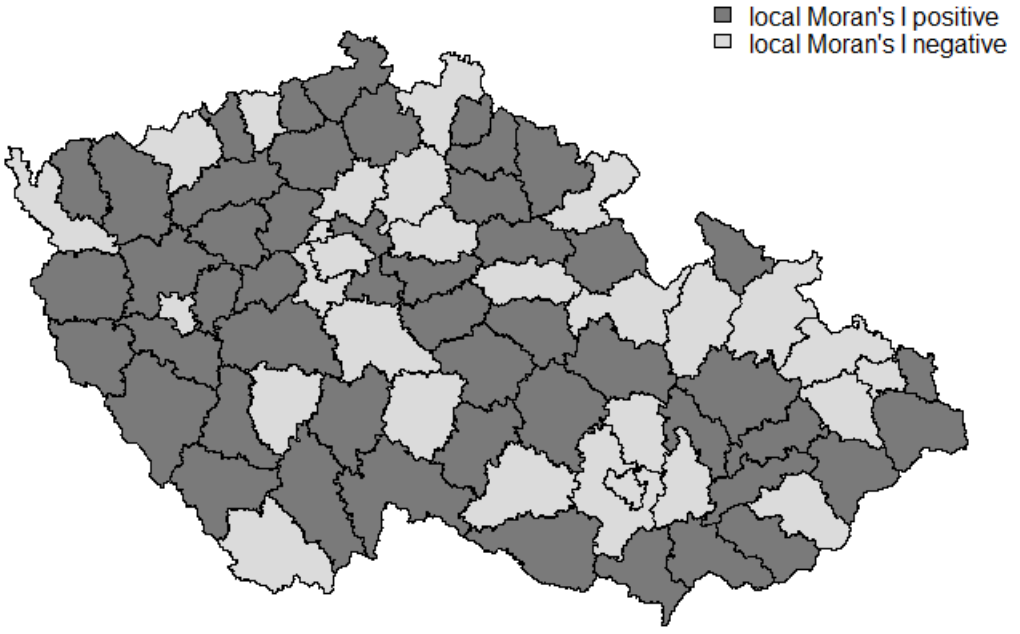

Figure 4.6: Local Moran's I - provision per capita

Source: Author based on shapefile by URRlab (2013).

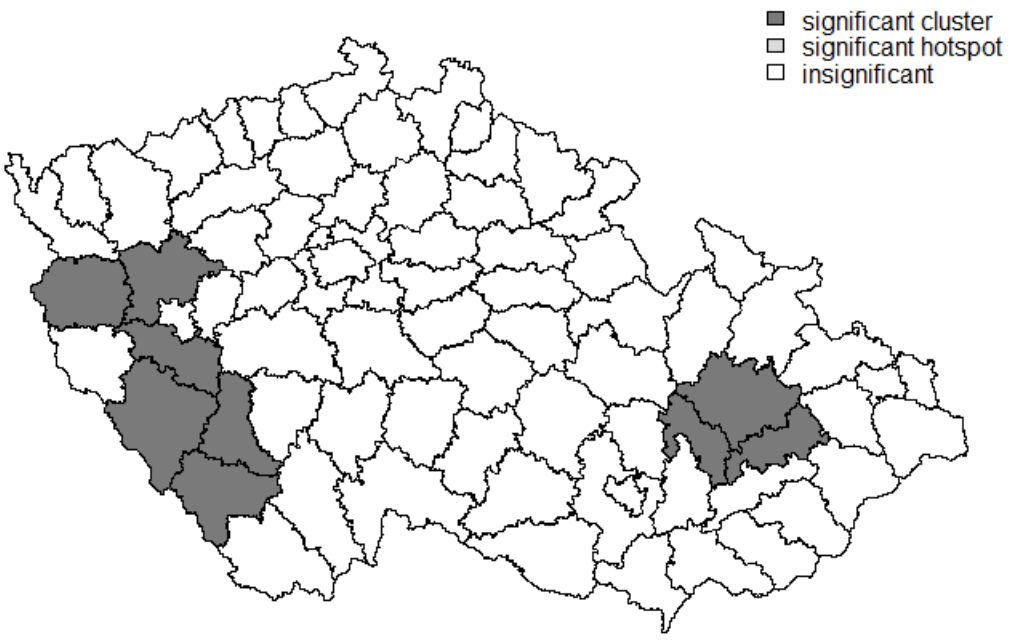

Figure 4.7: Clusters and hotspots in care provision per capita

Source: Author based on shapefile by URRlab (2013).

In order to get a better insight into the spatial autocorrelation structure, one can plot the value of local Moran's I for provision per capita as well, which was done in Figure 4.6. It is visible in Figure 4.6 that compared to care utilization, there are many more negative values of local Moran's I in the case of care provision. This corresponds to the result of the global test. When exploring clustering and hotspots in provision per capita, it is observable in Figure 4.7 that there are no hotspots again, however, there are also only two clusters that are significant at 5% level (changing the significance level to 10% does not change the result). Based on the local relationships, it is apparent that the result of the global test is quite intuitive.

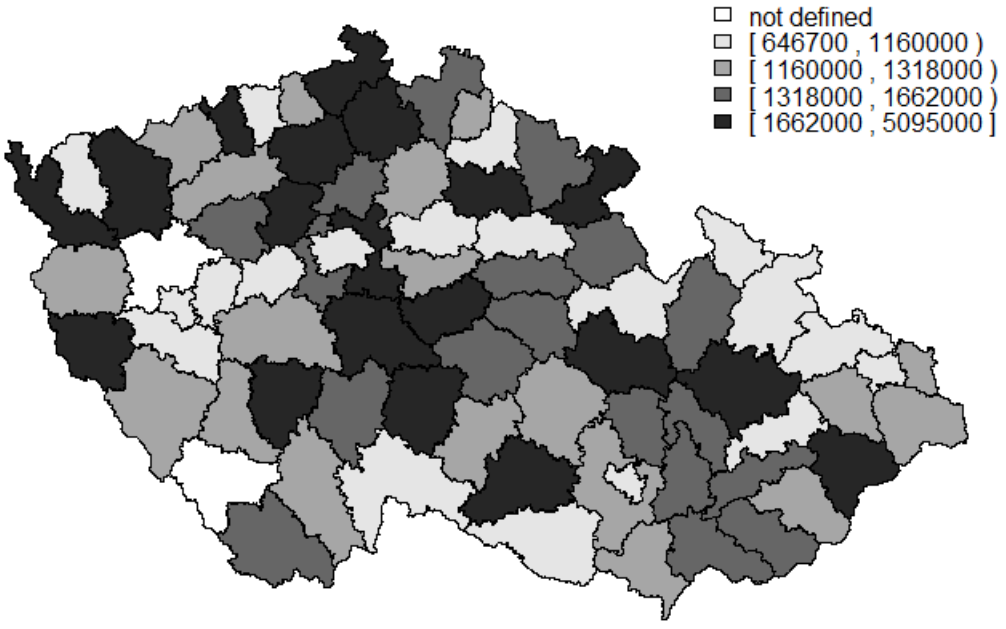

Figure 4.8: Points provided per ACI physician (map)

Source: Author based on shapefile by URRlab (2013).

Another interesting measure of care provision is the number of points for provision per physician ($ProvPP$). Figure 4.8 provides graphical representation of its values in a map. The variable $ProvPP$ is not defined in two districts because of zero number of ACI physicians (see Section 3.3 for details). In the map, it is very hard to find any visible correlation in the data. The visual impression is correct because the result of global Moran's I test (p-value of 0.51) points strongly towards no spatial autocorrelation in the variable as the null hypothesis cannot be rejected at any reasonable significance level.

The global relationship is, naturally, reflected by the local statistic as well. In Figure 4.9 one can see a high number of negative values of local Moran's I which

indicates that neighboring observations differ. Moreover, if one takes statistical significance into account, one finds out that there is only one district with a significantly positive value and no district with a significantly negative one (see Figure 4.10). Hence there is no spatial autocorrelation in $ProvPP$ whatsoever.

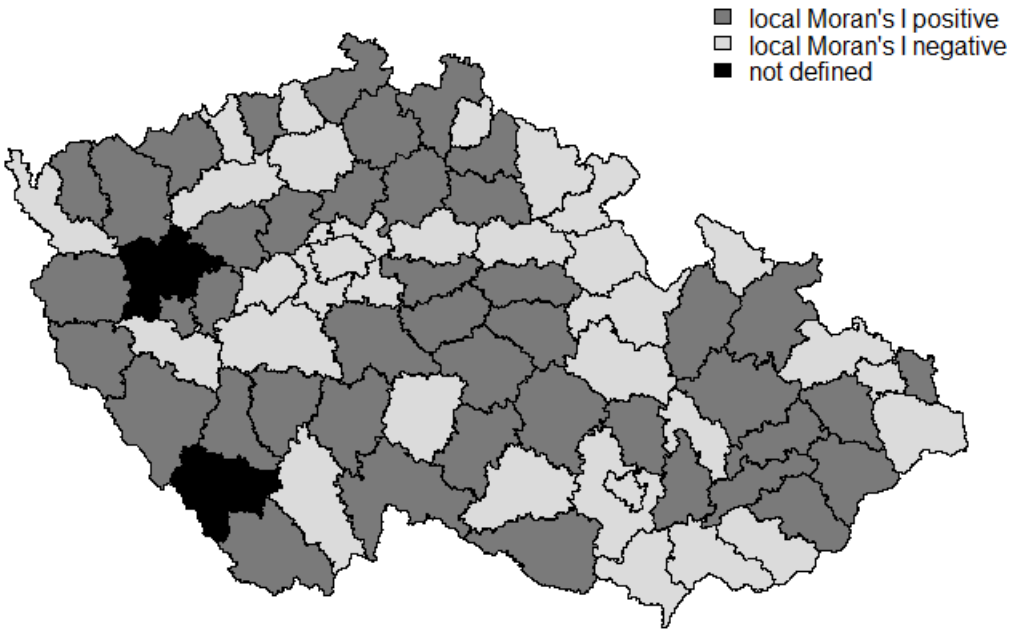

Figure 4.9: Local Moran's I - provision per ACI physician

Source: Author based on shapefile by URRlab (2013).

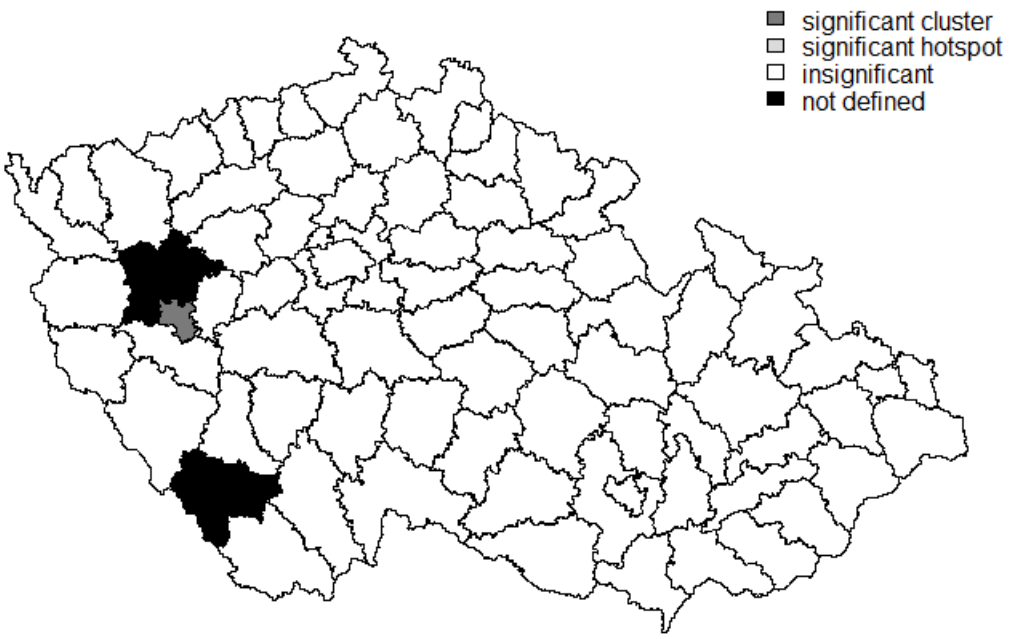

Figure 4.10: Clusters and hotspots in provision per ACI physician

Source: Author based on shapefile by URRlab (2013).

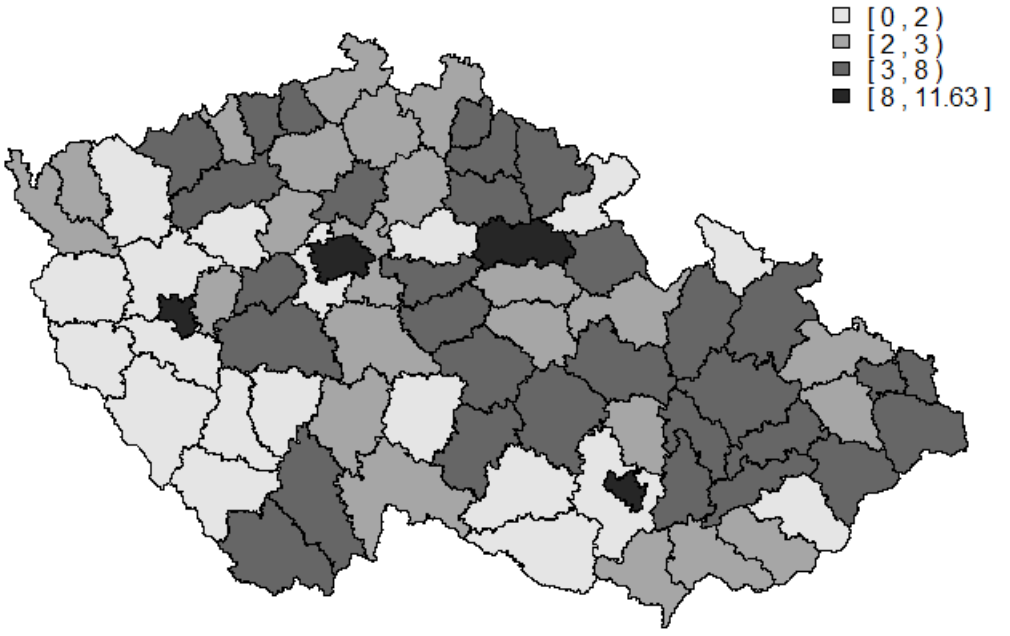

Figure 4.11: ACI physicians per capita (map)

Source: Author based on shapefile by URRlab (2013).

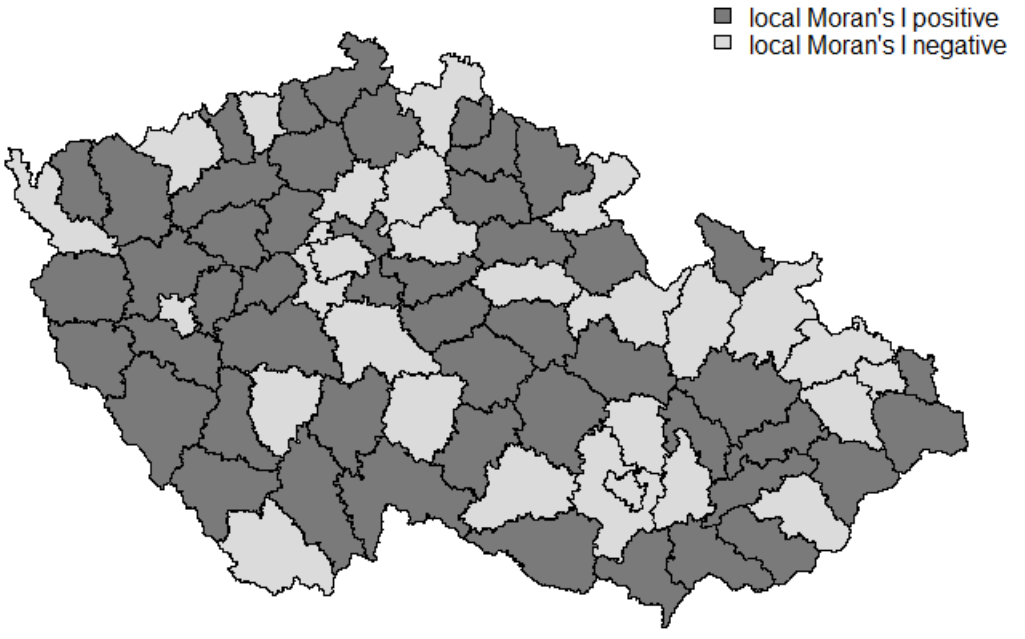

Figure 4.12: Local Moran's I - ACI physicians per capita

Source: Author based on shapefile by URRlab (2013).

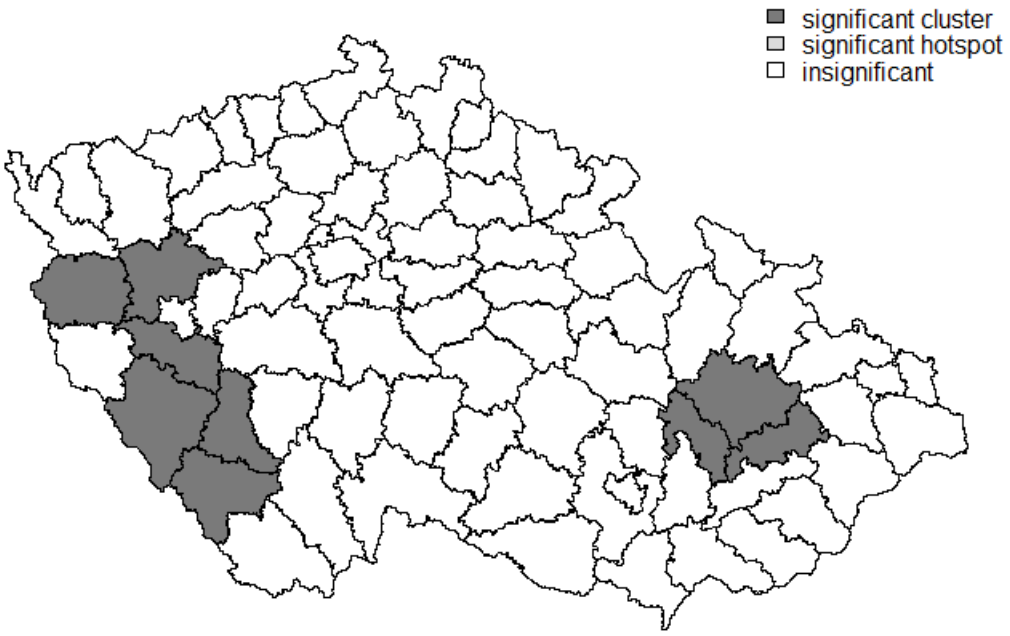

Figure 4.13: Clusters and hotspots in ACI physicians per capita

Source: Author based on shapefile by URRlab (2013).

It is also highly interesting to explore spatial relationships in the number of ACI physicians per capita in districts ($PhysPC$). The values of this variable are mapped in Figure 4.11, from which one can see that many times, districts with very different values are contiguous to each other. It is therefore intuitive that the global Moran's I test with p-value 0.66 clearly states that there is no spatial autocorrelation in the variable at any reasonable significance level. This is supported by the many negative values of local Moran's I in Figure 4.12. Although there are two significant clusters at 5% level visible in Figure 4.13, one can see from the figure that overall, there is not much spatial dependence. It is very interesting that the map in Figure 4.13 is identical to the one in Figure 4.7 showing care provision per capita. This is, nevertheless, not surprising as the tautological relationship between these two variables has already been explained in Section 3.3.

In order to make the analysis of spatial autocorrelation in variables complete, this section will be concluded by an overview of results of the global Moran's I test for the rest of the variables described in Section 3. The p-values of the test are provided in the second column of Table 4.1. Based on the p-values, one can decide about the presence of spatial autocorrelation in the variables. The significance at 5% level is denoted in the third column of Table 4.1. It is interesting but not unexpected that almost all of the variables are very significantly spatially autocorrelated. This is in accordance with one's intuition that environmental variables as well as socio-

economic characteristics should have similar values in contiguous districts. There are only two exceptions. The first one is the population density for which spatial autocorrelation is significant at 10% level only. The second one is the percentage of points provided by ACI physicians (out of total points provided by all specialties) for which no spatial autocorrelation cannot be rejected even at 30% level and is therefore concluded not to be spatially autocorrelated.

Table 4.1: Global Moran's I test of explanatory variables

Variable name	Global Moran's I test: p-value	Spatial autocorrelation significant at 5%?
Average age	$5.56 \cdot 10^{-3}$	yes
Index of ageing	$3.39 \cdot 10^{-3}$	yes
Unemployment rate	$5.29 \cdot 10^{-3}$	yes
Population density	$9.44 \cdot 10^{-2}$	no
Percentage provided by ACI	$3.89 \cdot 10^{-1}$	no
Arable land	$5.29 \cdot 10^{-8}$	yes
Gardens and orchards	$4.43 \cdot 10^{-9}$	yes
Permanent grassland	$3.10 \cdot 10^{-11}$	yes
Hop gardens and vineyards	$2.26 \cdot 10^{-6}$	yes
Forest land	$9.24 \cdot 10^{-5}$	yes
Water body areas	$1.49 \cdot 10^{-5}$	yes
Built-up areas and yards	$4.31 \cdot 10^{-3}$	yes
Other areas	$5.61 \cdot 10^{-4}$	yes
Emissions: solids	$9.83 \cdot 10^{-6}$	yes
Emissions: sulphur dioxide	$4.44 \cdot 10^{-6}$	yes
Emissions: nitrogen oxides	$4.17 \cdot 10^{-5}$	yes
Emissions: carbon monoxide	$1.49 \cdot 10^{-9}$	yes
Protected areas	$1.32 \cdot 10^{-5}$	yes
Roads length	$5.51 \cdot 10^{-5}$	yes

Note: As $pctgACI$ is not defined for Prachatice district, the corresponding row and column of matrix W were dropped too when the test was performed.
Source: Author's computations.

In this section, it was shown that spatial relationships are very important in the study of care utilization as it is highly spatially autocorrelated. Moreover, the majority of its determinants exhibits spatial autocorrelation as well. The noticeable exception is the per capita number of ACI physicians, which is not spatially autocorrelated. Interestingly, care provision per capita and care provision per ACI physician also do not show spatial autocorrelation which points to the conclusion that the geographic structures of care provision and utilization are very different.

5 Applied Econometric Methods

This section explains the methodology used in the present thesis. First, the application of the analysis of principle components to environmental characteristics is described as this method will be used for decreasing the number of used variables. Second, the method of ordinary least squares is explained. Although this method may seem trivial to certain economists, it is widely used in health economics (e.g. by Shaw et al., 2005). Moreover, it is the first step and a benchmark for the estimation of models in spatial econometrics. Finally, the spatial econometrics methods used for estimation are described as the data exhibit spatial autocorrelation and thus a necessity to use spatial econometrics is likely to arise.

5.1 Principal Component Analysis of Environmental Variables

As already explained in Section 2.1, despite the fact that it is clear that environment has an impact on allergic diseases, it is, however, not clear what particular aspects of environment influence allergy most. Therefore 14 environmental variables described in detail in Section 3.6 were obtained for each Czech district.

When using these environmental characteristics, one must overcome the following problems:

(i) It is not clear which of these variables influence allergic diseases. The majority of studies on allergy points to environment in general when naming causes of allergy. The only mechanism, on which one could base the decision about what variables to use, is intuition.

(ii) Some of the variables are highly correlated. When computing correlations among the variables, one discovers that out of the 91 correlation coefficients 7% are higher than 0.7 and 4.5% are higher than 0.8. Therefore, one would have to choose among some of the variables as to which ones to use because using two highly correlated ones might drive regression coefficients insignificant. (Wooldridge, 2009)

(iii) There are only 77 observations in the dataset and therefore using all 14 of these variables would greatly reduce degrees of freedom. Again, this might lead to loss of significance. (Wooldridge, 2009)

Principal Component Analysis (PCA) is used to overcome these issues. As described by Jolliffe (2002), PCA can be used to decrease the number of variables used while maintaining a high portion of variance in only a few orthogonal new variables. This method is, therefore, very useful when applied to the environmental dataset considering the three issues related to it mentioned previously as it will reduce the number of used variables and remove correlation.

Jolliffe (2002) describes in detail the general method of PCA based on which the application to the environmental dataset follows. PCA is an orthogonal linear transformation of data which transforms the original data into a new coordinate system. The property of interest of the new coordinate system is that the greatest variance is projected to the first coordinate, also called the first Principal Component (PC). The second largest portion of variance is projected to the second coordinate (or the second PC) etc.

When deriving the first PC mathematically, consider a matrix including all environmental characteristics in columns, say $Envir = [envir_1 \quad \cdots \quad envir_{14}]$. The PCA is performed by first looking for a vector α_1 which maximizes variance of $\alpha_1^{'}Envir$ which is in fact a linear combination of the original variables in $Envir$. The next step is to look for a vector α_2 which maximizes variance of $\alpha_2^{'}Envir$ while keeping $cor(\alpha_1^{'}Envir\,, \alpha_2^{'}Envir) = 0$. If one continues in this procedure, always finding a variance-maximizing linear function $\alpha_k^{'}Envir$ which is uncorrelated with $\alpha_1^{'}Envir\,, \ldots, \alpha_{k-1}^{'}Envir$, one is able to find 14 such functions as there are 14 original variables in $Envir$. These linear functions of the original variables are their principal components.

The `R` software was used for calculation of the principal components. The author used function `prcomp` from package `stats`. This function uses the singular value decomposition method to find the PCs as this method is numerically very accurate. Because the original variables have very different scaling (some of them are in percentages, some are values in tons or meters), scaling of the centered original variables was performed so that they had unit variance. The scaled and centered variables are then used for computation of the PCs.

As one of the reasons for using PCs is the decrease of the number of explanatory variables, one must make a choice about which of the PCs to use. As

each PC contains a higher proportion of variance in the original data than the following one, it is clear that the decision will be about how many of the first PCs one should use. Taking advantage of previous data scaling and centering in this application, one can use the Kaiser's rule which states that only the PCs with variance greater than 1 should be used. This is due to the fact that if one omitted this rule, one would include PCs which contain less variance than the original variables. Moreover, applying this rule ensures the usage of one and only one PC for each group if groups of correlated variables were present as all other PCs associated with that group would not satisfy the Kaiser's rule.

After determining what PCs describe the original data best, one can use those PCs in further estimation instead of the original variables.

5.2 Ordinary Least Squares for Cross-Sectional Data

The method of Ordinary Least Squares (OLS) for multiple linear regression models is described by Wooldridge (2009), on whom Section 5.2 is based (unless stated otherwise). The general multiple linear regression model for n observations with k explanatory variables can be written as

$$y = X\beta + u$$

where y is an $n \times 1$ vector containing the independent variable, X is an $n \times (k+1)$ matrix containing ones in the first column and the k explanatory variables in other columns, u is the $n \times 1$ vector of disturbances, and $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$ is the vector of regression parameters with β_0 representing the intercept and $\beta_1, \dots, \beta_k$ the coefficients on explanatory variables.

The OLS method finds estimates $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k)'$ of the regression parameters simultaneously, so that the sum of squared residuals

$$(y - X\hat{\beta})'(y - X\hat{\beta})$$

is minimized. The solution of this minimization problem is

$$\hat{\beta} = (X'X)^{-1}X'y$$

which is the OLS estimator of β.

The properties of OLS estimators rely on a number of assumptions:

(i) linearity in parameters,

(ii) random sampling,

*(iii)*no perfect collinearity among independent variables,

*(iv)*zero conditional mean: $\mathrm{E}(u|X) = 0$,

(v) homoskedasticity: $\mathrm{Var}(u|X) = \sigma^2$,

*(vi)*normality: u is independent of X and $u \sim \mathrm{N}(0, \sigma^2)$.

Under assumptions *(i)-(iv)*, the OLS estimator is unbiased and consistent. Assumptions *(i)-(v)* allow for computation of unbiased variances and standard errors of $\hat{\beta}_i$ for $i = 1, \dots, k$. Under assumptions *(i)-(v)*, OLS is the best linear unbiased estimator of $\beta = (\beta_0, \beta_1, \dots, \beta_k)'$, which means that its variance is smallest of all linear unbiased estimators. Under assumptions *(i)-(vi)*, the standardized estimators $(\hat{\beta}_i - \beta_i)/se(\hat{\beta}_i)$ follow t_{n-k-1} distribution, which allows to test linear hypotheses about regression coefficients using t test and F test. In case assumption *(vi)* does not hold and exact inference based on t and F statistics is not possible, one can still rely on asymptotic normality thanks to the central limit theorem in large enough samples. In this case, one can rely on the approximate t and F distributions of t and F statistics. However, there is unfortunately no rule on how large the sample must be for the approximation to work, although many econometricians agree that the central limit theorem can be applied if at least 30 observations are available.

The OLS assumptions need to be tested for, so that one can assess the OLS properties. Model misspecification is tested for by the RESET test, which assumes under its null hypothesis that no powers of used variables are missing in the model. Homoskedasticity is tested for by the Breusch-Pagan test and the White's test. The null hypothesis of both tests is homoskedasticity. The White's test tests against heteroskedasticity of general form, which has an advantage that it does not overlook any possible type of heteroskedasticity, however, the test might also detect a specification error instead of heteroskedasticity. The studentized version of the Breusch-Pagan test by Koenker (Greene, 2012) is used in order to avoid argued reliance on normality. The alternative of the Breusch-Pagan test is the error variance being a multiplicative function of used variables, which makes the test less general. Normality is tested using the Shapiro-Wilk test and the Jarque-Bera test (Thode, 2002). The null hypothesis of the Shapiro-Wilk test of a variable is that the variable has skewness and kurtosis of normal distribution (the alternative being that it has not).

The goodness of fit of an OLS model is measured by R^2, which quantifies what proportion of sample variation in the dependent variable is explained by the independent variables in the model. This measure, however, can be misleading when used for decisions about the best fitting model as it will always increase when an additional variable is added to the model. "Everything else being equal, simpler models are better" (Wooldridge, 2009, p. 202) and therefore adjusted R^2, which penalizes for the number of used independent variables, is a better measure of goodness of fit when making a choice among models.

Although the OLS method provides best linear unbiased estimators when its assumptions are satisfied and is thus useful in many applications despite its simplicity, spatial dependence in the data might occur which makes it necessary to use methods of spatial econometrics instead.

5.3 Spatial Econometrics Models for Areal Data

The description of spatial econometrics models and estimation methods in Section 5.3 and its subsections is (unless stated otherwise) based on LeSage & Pace (2009) who explain these in great detail.

In cross-sectional data, the usual assumption is that the observations of variables are independent. However, if one's data are of spatial character, the analysis might result in the emergence of spatial dependence[8]. In that case, one is dealing with a simultaneous data generating process where the value of a variable in one area depends on the values in neighboring areas and vice-versa. In this work, the areas are the 77 Czech districts, whose neighboring relationships are described in Section 3.1, where the binary contiguity matrix W^* and the row-standardized contiguity matrix W are constructed. Bivand et al. (2008) suggest taking the approach of a binary contiguity matrix whenever one can assume little about the spatial process (i.e. one should not set the weights in W^* arbitrarily different from zero and one without an objective reason).

The spatial dependence of observations of individual variables (which is the case of majority of used variables as described in Section 4) does not necessarily imply the use of spatial econometrics. For example if there are two variables x and y which are both spatially autocorrelated, one might still be able to estimate their

[8] The terms spatial data, spatial dependence, spatial lag, and spatial autocorrelation were explained in Section 4.

relationship by simple OLS because it might be the case that the spatial autocorrelation in y is driven by the influence of x on y. Therefore it is not true that spatial models have to be used for any kind of areal data. Nevertheless, omitted variables often occur in spatial modeling because variables describing latent influences are often unlikely to be readily available, which makes the use of spatial models necessary. On the other hand, the spatial autocorrelation in y can also be caused by a spillover effect which can be captured by spatial models as well.

There are several types of spatial models, each accounting for different spatial effects. Before moving to the description of these models, let us denote n the number of observations, y an $n \times 1$ vector containing the dependent variable, X the $n \times k$ matrix containing the k independent variables, W the $n \times n$ row-standardized contiguity matrix, ε the $n \times 1$ vector of disturbances, ι_n an $n \times 1$ vector of ones, I_n the $n \times n$ identity matrix, and 0 an $n \times 1$ vector of zeros.

5.3.1 Spatial Autoregressive Model and Spatial Durbin Model

When combining a conventional regression model with spatial autoregressive structure, one obtains the spatial autoregressive model (SAR) of the form

$$y = \rho W y + \alpha \iota_n + X\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

The parameters to be estimated in this model are the usual ones α, β, and σ and the new spatial lag parameter ρ. This model employs, apart from the independent variables, also the spatial lag of the dependent variable.

The data-generating process of the SAR model is

$$y = (I_n - \rho W)^{-1}(\alpha \iota_n + X\beta) + (I_n - \rho W)^{-1}\varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

The spatial Durbin model (SDM) is a natural extension of the SAR which includes spatial lags of independent variables:

$$y = \rho W y + \alpha \iota_n + X\beta + WX\gamma + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

The additional parameters to be estimated here are in γ. The data-generating process of the SDM is

$$y = (I_n - \rho W)^{-1}(\alpha \iota_n + X\beta + WX\gamma + \varepsilon), \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

The estimation of neither SAR nor SDM can be done by OLS because it would lead to bias and inconsistence of parameter estimators. Therefore Maximum Likelihood Estimation (MLE) is used as this method is consistent for both the SAR

and the SDM. The likelihood functions of SAR and SDM coincide because if we denote $Z = [\iota_n \;\; X]$ for SAR and $Z = [\iota_n \;\; X \;\; WX]$ for SDM, we can describe both models by

$$y = \rho W y + Z\delta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n)$$

together with their data generating process

$$y = (I_n - \rho W)^{-1} Z\delta + (I_n - \rho W)^{-1}\varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n).$$

If one decided to maximize the full log-likelihood, one would have to solve the first order conditions for parameters δ, ρ, and σ^2. This is, however, not necessary as it is possible to concentrate the full likelihood with respect to parameters δ and σ^2 and solve a univariate optimization problem for ρ instead. Moreover, if one works with the concentrated log-likelihood, one arrives to the same estimates of the parameters that would be produced when maximizing the full log-likelihood. The log-likelihood function of SDM and SAR is of the form

$$\begin{aligned}
&\log(L) = -\frac{n}{2}\log(\pi\sigma^2) + \log|I_n - \rho W| - \frac{e'e}{2\sigma^2} \\
&e = y - \rho W y - Z\delta \\
&\rho \in (min(\omega)^{-1}, max(\omega)^{-1})
\end{aligned}$$

where $|\cdot|$ denotes matrix determinant and ω is an $n \times 1$ vector of eigenvalues of W. Please note that these eigenvalues are real for the contiguity matrix used in this work because W has the same eigenvalues as W^* which is a symmetric matrix (see Section 3.1 for details). The condition $\rho \in (min(\omega)^{-1}, max(\omega)^{-1})$ ensures that the variance-covariance matrix is positive definite. The concentrated log-likelihood then takes the form

$$\log\big(L(\rho)\big) = \kappa + \log|I_n - \rho W| - \frac{n}{2}\log\big(S(\rho)\big)$$

where κ is a constant independent of ρ and

$$\begin{aligned}
&S(\rho) = e(\rho)'e(\rho) \\
&e(\rho) = e_0 - \rho e_d \\
&e_0 = y - Z\delta_0 \\
&e_d = Wy - Z\delta_d \\
&\delta_0 = (Z'Z)^{-1}Z'y \\
&\delta_d = (Z'Z)^{-1}Z'Wy.
\end{aligned}$$

The MLE method is then applied to the concentrated log-likelihood and the optimum value of ρ is then its maximum likelihood estimate $\hat{\rho}$. The computation of other parameters of the model is based on the value of $\hat{\rho}$. Thus, the estimates of coefficients $\hat{\delta}$, the noise-variance parameter $\hat{\sigma}^2$, and the disturbance variance-covariance matrix $\widehat{\Omega}$ can be expressed as

$$\hat{\delta} = \delta_0 - \hat{\rho}\delta_d$$
$$\hat{\sigma}^2 = n^{-1}S(\hat{\rho})$$
$$\widehat{\Omega} = \hat{\sigma}^2[(I_n - \hat{\rho}W)'(I_n - \hat{\rho}W)]^{-1}$$

respectively.

In practice, the MLE of ρ can actually be followed by several types of univariate optimization techniques for obtaining $\hat{\delta}$, $\hat{\sigma}^2$, and $\widehat{\Omega}$. In this work, the estimation of SAR and SDM is performed using the `R` function `lagsarlm` from `spdep` package, which uses generalized least squares method to solve these univariate problems.

5.3.2 Interpretation of the SAR Model

Even if all the estimates of parameters of the SAR model are obtained, it is still not possible to interpret the model. This issue stems from the complex spatial structure of the model which causes that a change of value of an explanatory variable in any district will affect not only the dependent variable in that district but also the dependent variable in possibly all other districts. Therefore, the impact of any such change will be different for different districts as these have different structure of neighbors and the usual interpretation of coefficients as partial derivatives no longer holds.

In general, the total impact of a change in observation i of an explanatory variable can be divided into two parts. The direct impact of this change is the change in observation i of the dependent variable. There is, however, also the indirect impact which includes the effect of feedback loops in the dependent variable where "observation i affects observation j and observation j also affects observation i as well as longer paths which might go from observation i to j to k and back to i" (LeSage & Pace, 2009, p. 35). To put this in the ACI context, let us assume that the number of physicians has a positive relationship with care utilization. Then the direct effect of additional physicians in a district would be that care utilization in that district increased. The indirect effect of this change would then be the increase in care utilization in contiguous districts (and districts contiguous to contiguous districts etc.) that would be reflected through the neighboring relationships back to the

original district. The magnitude of the indirect impact depends on the neighboring structure of districts captured in matrix W, the parameter ρ measuring the spatial dependence in the dependent variable, and the parameters in β that measure the impact of explanatory variables on the dependent variable.

As the impact structure is very complicated because impacts differ depending on the observation in which a change is made and also on the observation for which we measure the impact, summary impact measures averaged over all observations are used for interpretation of the SAR. The average direct impact, average total impact, and average indirect impact of changes in the r-th independent variable on the dependent variable can be computed as

$$\begin{aligned} ADI(r) &= n^{-1}\mathrm{tr}\big(S_r(W)\big) \\ ATI(r) &= n^{-1}\iota_n' S_r(W)\iota_n \\ AII(r) &= ATI(r) - ADI(r) \end{aligned}$$

respectively, where $\mathrm{tr}(\cdot)$ denotes matrix trace and

$$\begin{aligned} S_r(W) &= V(W)I_n\beta_r \\ V(W) &= (I_n - \rho W)^{-1} = I_n + \rho W + \rho^2 W^2 + \rho^3 W^3 + \cdots . \end{aligned}$$

Due to the complexity of computation of $V(W)$ which includes an inversion of the $n \times n$ matrix $(I_n - \rho W)$, it is often computationally inefficient to compute the impacts exactly. However, with present-day computing power, this inversion including a 77×77 sparse matrix W used in this study does not cause any problems and therefore there is no need for any approximation techniques in this particular case.

When computed, the impacts for the SAR model can then be interpreted in the usual partial-derivative manner in which one is used to interpret coefficients from models without a spatial structure.

However, if one wishes to draw inferences about the impacts, their statistical significance is crucial. Therefore, one needs to construct an empirical distribution of parameters α, β, ρ, and σ. In this work, the empirical distribution is "constructed using a large number of simulated parameters drawn from the multivariate normal distribution of the parameters implied by the maximum likelihood estimates" (LeSage & Pace, 2009, p. 39).

In this work, the computation of impacts as well as their empirical distributions is performed using the `R` function `impacts` from `spdep` package, which uses the estimation procedure described in this section.

5.3.3 Spatial Error Model

The spatial lag might also arise in the disturbances process, leading to the spatial error model (SEM) of the form

$$y = \alpha\iota_n + X\beta + u$$
$$u = \rho Wu + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2 I_n)$$

where there are no spatial lags of neither the dependent nor independent variables. On the other hand, the error in this model is spatially autocorrelated. If the data follow the SEM, OLS can still be used for estimation of the model parameters (except for ρ) as their estimates will be unbiased, however, inefficient. Moreover, no consistent estimate of ρ can be obtained by OLS.

The estimation techniques of the SEM are not discussed here as the model is only tested for and the spatial autocorrelation in the error proves insignificant. SAC and SARMA models are extensions of the SEM combined with SAR, however, as they are not used in this work (based on insignificance of the spatial lag of the error), it is redundant to explain them in detail. However, if the reader is interested in how the SEM model is estimated or in details on SAC and SARMA models, LeSage & Pace (2009, p. 50-54) provide detailed information on the subject.

5.3.4 Spatial Econometrics Tests and Evaluation of Fit

Spatial econometrics, as any other type of econometrics, presents many tests and statistics of several kinds that help one to test hypotheses and find the best-fitting model. This subsection describes the tests and statistics used in this work.

Once the basic OLS model is estimated, it is necessary to find out whether a spatial lag of the dependent variable or of the error term is missing in the model. A version of the Moran's I test for linear regression models can be performed on OLS regression residuals in order to find out about residual spatial autocorrelation. As in the single-variable version of the test, its null hypothesis is no spatial autocorrelation, while the alternative can have two forms: either positive or no spatial autocorrelation. A more complex testing framework is provided by the set of four Lagrange Multiplier (LM) tests. The two basic tests test for an omitted spatially lagged dependent variable and spatially autocorrelated residuals respectively. The lacks of these effects are the tests' null hypotheses, while their presences are represented by the alternatives. The robust pair of tests test each for one of the effects in the possible presence of the other. Their hypotheses are analogous. (Anselin, 1988; Bivand et al., 2008)

The most intuitive hypothesis to be tested in the SAR model is probably the significance of the ρ coefficient on the spatially lagged dependent variable. There are three asymptotically equivalent tests that are commonly used: the LM test, the Likelihood Ratio (LR) test, and the Wald test. All three tests test the $H_0: \rho = 0$ against $H_1: \rho \neq 0$. The difference between the tests is in their measures of the distance between an unrestricted estimate and an estimate satisfying H_0. The Wald test is sometimes referred to as the asymptotic t-test. The LR test is based on the difference between the log-likelihoods of models under H_0 and H_1. The LM test is based on optimization of a lagrangian function in the log-likelihood. The testing statistics of all three tests follow χ^2 distribution with 1 degree of freedom (for this particular null hypothesis). In finite samples, the values of the three testing statistics can be ordered in the following manner: $Wald \geq LR \geq LM$. (Anselin, 1988)

Testing the significance of individual explanatory variables in X in the SAR model is most commonly performed by the Wald test using the property that "the square root of the Wald test corresponds to a standard normal variate" (Anselin, 1988, p. 67) and its asymptotic equivalence to the t-variate. Joint significance of several explanatory variables in the SAR model is performed by the LR test with the restricted model under H_0 and the unrestricted one under H_1. The number of degrees of freedom is equivalent to the number of restrictions and the testing statistic follows χ^2 distribution. (Anselin, 1988; LeSage & Pace, 2009)

Residual autocorrelation in the error term might occur in the SAR model causing a necessity to estimate the SAC or SARMA model instead. In order to test for this residual autocorrelation, an LM test is used. The null hypothesis of this test is no residual spatial autocorrelation, whereas the alternative is the nulls negation. The testing statistic is again χ^2 distributed with 1 degree of freedom. (Anselin, 1988)

Homoskedasticity of residuals in the SAR model is tested by the spatial version of the studentized Breusch-Pagan test. The test is performed by regressing the SAR residuals on the right-hand-side variables of the model, with the null hypothesis being homoskedasticity and the alternative being heteroskedasticity as a function of right-hand-side variables of the SAR model. The testing statistic follows χ^2 distribution. (Anselin, 1988)

The normality assumption needed for MLE is necessary to be tested as well. The test of normality of the SAR residuals can be performed by the usual Jarque-Bera or Shapiro-Wilk tests, where normality is the null hypothesis of both. (Anselin, 1988; LeSage & Pace, 2009; Thode, 2002)

There are several measures that evaluate the fit of the SAR (as well as of the SDM) model. The maximized value of the log-likelihood can be used for assessment of the model fit. A better fitting model has a higher maximized value of the log-likelihood. However, if a variable is added to a model, the log-likelihood will always increase which makes it an imperfect measure according to which one can make decisions about what model fits the data best. This problem is solved when the Akaike Information Criteorion (AIC) is used for model comparison instead. The AIC explicitly takes into account the trade-off between parsimony and fit. The AIC measures the closeness of the assumed model to an unknown true model. The AIC incorporates a measure of likelihood as well as the number of estimated parameters. Therefore, "the AIC corrects or penalizes the assessment of goodness-of-fit given by the maximized likelihood by a factor which reflects the number of parameters" (Anselin, 1988, p. 246). AIC is therefore considered an appropriate measure that can be used for model comparison, while the AIC is lower for better-fitting models. (Anselin, 1988; Bivand et al., 2008)

The MLE method by which the SAR is estimated makes the standard R^2 invalid. Nevertheless one can use the so-called Nagelkerke pseudo-R^2 which takes the form of squared correlation between predicted and observed values. The Nagelkerke pseudo-R^2 keeps the usual interpretation of proportion of explained variation. In particular, it holds for the Nagelkerke pseudo-R^2 that $1 - R_N^2$ can be interpreted as the proportion of unexplained variation. Interestingly, the R_N^2 is consistent with the classical R^2 as when applied to a linear regression model it yields the same result. (Anselin, 1988; Nagelkerke, 1991)

When making the decision whether the SAR or SDM model should be used, one can use both the LR test and the AIC. The LR test tests the unrestricted SDM against the restriction that the spatially lagged independent variables are jointly insignificant (which is in fact the SAR model that nests in the SDM model). The testing statistic is again χ^2 distributed with the number of degrees of freedom equivalent to the number of spatially lagged independent variables in the SDM model (which is equal to the number of independent variables in the SAR model). The AIC on the other hand provides a measure of fit corrected for the number of used variables like for the comparison of any two models. (Anselin, 1988; LeSage & Pace, 2009)

6 Model Search and Results

This section describes the search for models for care utilization and provision and provides results of the best-fitting models found. First, the results of PCA of environmental variables need to be reported as the resulting PCs serve as explanatory variables. Second, a model for care utilization is developed and its results are provided and interpreted. Finally, modeling of care provision is explained. In both cases, OLS estimation is used as a necessary first step. Only then the best-fitting OLS model should be tested for residual spatial effects and a spatial model may be applied instead.

6.1 Environmental PCA

Based on the PCA technique described in Section 5.1, PCs for environmental variables were computed in order to be used in further econometric modeling.

Table 6.1: Proportion of variance explained by PCs

PC	Proportion of Variance	Cumulative Proportion
PC_1	0.33420	0.33420
PC_2	0.23560	0.56980
PC_3	0.10080	0.67060
PC_4	0.09500	0.76560
PC_5	0.06919	0.83475
PC_6	0.05930	0.89410
PC_7	0.04806	0.94212
PC_8	0.02055	0.96267
PC_9	0.01875	0.98142
PC_{10}	0.01006	0.99148
PC_{11}	0.00496	0.99644
PC_{12}	0.00286	0.99930
PC_{13}	0.00070	1.00000
PC_{14}	0.00000	1.00000

Source: Author's computations.

Table 6.1 shows the proportion of variance of the original environmental variables explained by PCs. The second column of the table provides the proportion of variance each PC explains individually. Indeed, the highest proportion is explained by the first component and then the proportion decreases, leaving practically no

proportion of variance left to the last component. In the third column, the cumulative proportion of variance explained by a component and all previous components together can be viewed. It is worth noticing that the first two PCs alone explain 57% of variance and that 89% of variance is explained using only six PCs.

Based on the cumulative proportion of variance explained, it is even clearer that only a few PCs will suffice when explaining the variation contained in the environmental variables. Thus, the Kaiser's rule is applied to the PCs so that only the components with variance greater than 1 are used. The variances can be found in Table 6.2 and expectedly the first component has largest variance and the variance then decreases. Based on the Kaiser's rule, only the first four PCs are to be used as those are the only ones with variance greater than 1. Based on Table 6.1, one can see that the first four components explain approximately 77% of variance in the original variables.

Table 6.2: Variance of PCs

PC	Variance
PC_1	4.678682
PC_2	3.298012
PC_3	1.411081
PC_4	1.330060
PC_5	0.968725
PC_6	0.830239
PC_7	0.672896
PC_8	0.287701
PC_9	0.262448
PC_{10}	0.140818
PC_{11}	0.069504
PC_{12}	0.039975
PC_{13}	0.009855
PC_{14}	0.000003

Source: Author's computations.

Being linear combinations of the original variables, PCs are often very difficult to interpret. As suggested by Jolliffe (2002), the interpretation of PCs should be based on their correlation with the original variables from which they were constructed. Table 6.3 therefore provides these correlations. When exploring the correlations, one can find an approximate interpretation of the components. PC_1 represents emissions as it is highly positively correlated with the emission measures, while other PCs are not. PC_2 could represent natural plant life as it is positively correlated mainly with forests, grassland and protected areas while being negatively correlated with arable land which is artificial plant life. PC_3 is very hard to interpret

as it is positively correlated mainly with built-up areas and yards, and gardens and orchards. It may represent residences, either cities or villages, as opposed to natural areas. The PC_4 is positively correlated with roads and negatively correlated with water body areas as well as hop gardens and vineyards. It may in a certain way describe pollution, nevertheless, the interpretation is not very precise. Fortunately, this is not a problem as it is not the purpose of the present thesis to identify the determinants of care utilization and provision. It is necessary to merely control for them in order to be able to explain as much variation in care utilization and provision as possible, for which the PCs are sufficient.

Table 6.3: Correlation of PCs and original environmental variables

	PC_1	PC_2	PC_3	PC_4
Arable land	0.20	- 0.90	- 0.31	0.05
Gardens and orchards	0.72	- 0.24	0.54	0.20
Permanent grassland	- 0.53	0.67	0.04	0.13
Hop gardens and vineyards	0.05	- 0.34	- 0.03	- 0.66
Forest land	- 0.61	0.69	0.05	0.03
Water body areas	0.28	0.03	- 0.03	- 0.56
Built-up areas and yards	0.75	- 0.06	0.62	0.08
Other areas	0.74	0.37	0.34	- 0.18
Emissions: Solids	0.81	0.43	- 0.22	0.14
Emissions: Sulphur dioxide	0.66	0.51	- 0.44	- 0.12
Emissions: Nitrogen oxides	0.75	0.49	- 0.35	- 0.12
Emissions: Carbon monoxide	0.61	0.34	- 0.13	0.19
Protected areas	- 0.43	0.57	0.22	0.09
Roads length	0.28	- 0.34	- 0.30	0.62

Source: Author's computations.

6.2 Care Utilization

This section describes modeling of care utilization using observations from 77 Czech districts. First, the dependent variable is chosen and potential explanatory variables are commented on. Next, the exogeneity of number of physicians in the model is explained. Finally, the model describing utilization of care is found, estimated, and its results are interpreted.

6.2.1 Used Dependent and Independent Variables

There is no particular theoretical model that could be used for modeling of care utilization. One, therefore, has to rely on the partial knowledge of the possible determinants of utilization of care and on a following econometric analysis to find the correct model. In order to include all theoretically important effect on care utilization,

one has to consider all variables that could affect the need for care by patients (age, environment, socio-economic situation) as well as variables describing capacities available for people to utilize care in (number of physicians in district and in surrounding districts).

The dependent variable of this analysis is the overall care utilization per capita in ACI ($UtilPC$). As this variable is positively skewed and not normally distributed according to the Shapiro-Wilk test (p-value 0.0129), the recommendation of Wooldridge (2009) to use its natural logarithm is applied. The variable $\log(UtilPC)$ follows normal distribution according to the Shapiro-Wilk test (p-value 0.1232) and also according to the Jarque-Bera test (p-value 0.2971).

The independent variables in the model will be chosen from the four PCs describing environment derived in Section 6.1 (PC_1, PC_2, PC_3, PC_4), the number of ACI physicians per capita ($PhysPC$, $NeighPhysPC$, $AvNeighPhysPC$), the age structure ($AvAge$, $OAindex$), the population density ($Pdens$), and the unemployment rate ($Unemp$). The functional form of these variables might be modified based on their distribution, significance, and model fit.

6.2.2 Notes on Exogeneity of Number of Physicians

The next step of the analysis is the use of the number of ACI physicians per capita in a district ($PhysPC$). First of all, it is necessary to assess whether this explanatory variable is truly an exogenous one as its endogeneity would mean a violation of assumption *(iv)* of OLS.

It is clearly not possible for the utilization of care to have an effect on the number of physicians within one year due to the reimbursement system in Czech health care described in Section 2.3 as physicians have long-term contracts with insurers. Nevertheless, some might argue that the past levels of care utilization have led to changes in the number of physicians in districts as physicians adapted to the needs of patients. The purpose of this section is to explore the possibility of such past adjustment.

In Section 4, it was shown that utilization per capita is strongly spatially autocorrelated. This means that utilization values tend to be similar in neighboring districts. This is not at all surprising as all used variables that are expected to have an impact on need of care by patients are also spatially autocorrelated. On the other hand, it has also been shown that the number of ACI physicians per capita is not spatially autocorrelated. On the contrary, it is visible from Figure 4.11 that the number of physicians per capita tends to be high mainly in districts representing big

cities: Prague, Plzeň-town, Brno-town. This holds despite the fact that the care utilization per capita does not tend to be higher in these areas than in surrounding areas as shown in Figure 4.2. This leads to a conclusion that the geographical structure of number of physicians does not correspond to the geographical structure of care utilization.

Considering also the fact that the issues covered by the branch of ACI are usually those that are very long-term and chronic and therefore unlikely to suffer from fast year-to-year changes in geographical structure (as depicted in Section 2.1), it may be seen that there is theoretically a lot of time for changes in the geographical structure of distribution of physicians for adjustment to care utilization. However, findings in Section 4 show that such adjustment did not take place and that physicians tend to group in large cities rather than to distribute themselves according to the utilization needs. Thus it can be concluded that the variable $PhysPC$ is exogenous. This conclusion is supported by the fact that the correlation between the number of ACI physicians per capita and utilization per capita is only 0.41.

This phenomenon, although it might seem contra intuitive to non-health economists, is quite common in Czech health care and it is considered rather problematic. Further issues describing long-term non-adjustment of the number of physicians in several branches of Czech health care to the population needs are discussed among many others by Rodriguez & Dolejší (2015), Šídlo & Tesárková (2009), and Šídlo (2010).

6.2.3 Model Search

Prior to actual model estimation, it is necessary to analyze what forms of explanatory variables seem suitable to be used. It is also useful to analyze the relationship with the dependent variable $\log(UtilPC)$ of groups of variables among which it is necessary to make a choice and use only one. The level of significance used for statistical tests is the standard 5%.

As the number of ACI physicians per capita $PhysPC$ is not normally distributed according to both the Jarque-Bera test and Shapiro-Wilk test (p-values $2.20 \cdot 10^{-16}$ and $2.62 \cdot 10^{-8}$ respectively) and it is positively skewed with skewness parameter 1.92, the recommendation of Wooldridge (2009) to use natural logarithm in such cases is applied. For the two observations where the number of ACI physicians per capita is zero, however, the logarithm does not exist. Therefore the standard approach described also by Wooldridge (2009) is followed and the variable $\log(PhysPC+1)$ is used for estimation instead. The advantage of this transformation

is that the new variable is positive whenever the original one is positive and it is zero when the original one is zero. This transformation, although it does not ensure normal distribution, moves the variable $\log(PhysPC+1)$ much closer to normal according to the Jarque-Bera test and the Shapiro-Wilk test with p-values $1.62 \cdot 10^{-3}$ and $3.30 \cdot 10^{-3}$ respectively.

While considering the impact of physicians in neighboring districts on care utilization, two variables are available: the number of ACI physicians per capita in neighboring districts $NeighPhysPC$ and the average number of ACI physicians per capita in neighboring districts $AvNeighPhysPC$. The variable $NeighPhysPC$ should also be used in log form as it is positively skewed (skewness parameter 0.77) and not normal (p-value of the Jarque-Bera test is $1.80 \cdot 10^{-2}$ and it is $3.50 \cdot 10^{-3}$ for the Shapiro-Wilk test). Normality is achieved in $\log(NeighPhysPC)$ as the p-values of the Jarque-Bera and Shapiro-Wilk test are 0.79 and 0.69 respectively. The correlation with $\log(UtilPC)$ changes only slightly from 0.3088 for the original variable to 0.2985 for the logged one. The variable $AvNeighPhysPC$ is normally distributed (Jarque-Bera test p-value 0.55, Shapiro-Wilk test p-value 0.77), however, the also normal log version increases correlation with $\log(UtilPC)$ from 0.3082 to 0.3207. Thus, $\log(NeighPhysPC)$, $AvNeighPhysPC$, or $\log(AvNeighPhysPC)$ seem to be good candidates to describe the neighboring ACI physician structure.

It is also necessary to incorporate the age structure into the models, which can be done by using either the average age $AvAge$ or the index of ageing $OAindex$. Neither of these variables is normally distributed according to both the Jarque-Bera and Shapiro-Wilk tests, although the use of natural logarithm does not make an improvement in normality. On the contrary, $\log(OAindex)$ is even further from normal distribution. However, using logs might bring improvements in terms of fit as magnitude of correlation with $\log(UtilPC)$ increases slightly when implementing log for $AvAge$ from -0.1382 to -0.1406. In the case of $OAindex$, the original correlation is -0.0765 while the correlation with the logged variable is -0.0969. According to the correlation coefficients, $AvAge$ or $\log(AvAge)$ seem to be best candidates for describing the age structure.

The use of population density in the model is problematic as $Pdens$ is highly correlated with $\log(PhysPC+1)$ with correlation coefficient 0.4804, which might lead to insignificance if both are included according to Wooldridge (2009). The variable is far from normally distributed (p-value $2.20 \cdot 10^{-16}$ for both the Jarque-Bera and Shapiro-Wilk test) and applying natural log helps neither with normality improvement which is only marginal, nor with lowering the correlation with

$\log(PhysPC{+}1)$ as it actually increases to 0.5637 when $\log(Pdens)$ is used. This is in accord with the finding that physicians group in larger cities as $cor(PhysPC, Pdens)$ is as high as 0.6205.

The economic situation is proxied by the unemployment rate $Unemp$. The variable is normally distributed (p-value 0.34 for the Jarque-Bera and 0.25 for the Shapiro-Wilk test), however, its correlation with $\log(UtilPC)$ is as low as -0.0043 and thus it is not likely to explain $\log(UtilPC)$ well. Using its logarithm (which is also normally distributed) slightly increases the magnitude of correlation to -0.0169, nevertheless, it still remains very low.

The four principal components PC_1, PC_2, PC_3, PC_4 will be used for estimation as long as they are jointly significant.

Based on the groups of explanatory variables described above, 108 models including one variable from each group were estimated by OLS, both including and excluding population density and unemployment rate. Particularly, the groups from which variables were chosen can be summarized as:

(i) $AvAge$, $\log(AvAge)$, $OAindex$, $\log(OAindex)$,

(ii) $\log(NeighPhysPC)$, $\log(AvNeighPhysPC)$, $AvNeighPhysPC$,

(iii) $\log(Pdens)$, $Pdens$, or none,

(iv) $\log(Unemp)$, $Unemp$, or none,

(v) the group of four PCs,

(vi) $\log(PhysPC{+}1)$

and the choice of variables in the groups was based on their above-mentioned distributions and correlation with $\log(UtilPC)$.

The best OLS model was chosen based on adjusted R^2 and significance of all used variables. The OLS model that was found to best describe the data is

$$\log(UtilPC) = \beta_0 + \sum_{k=1}^{4} \beta_k PC_k + \beta_5 \log(PhysPC{+}1) + \beta_6 \log(AvAge) + \beta_7 \log(AvNeighPhysPC) + u$$

where u is the disturbance term and β_0, ... , β_7 are the parameters to be estimated. The OLS coefficient estimates together with p-values of the t statistics are provided in Table 6.4. The regression R^2 is 0.4887 and the adjusted R^2 is 0.4368. The p-value of the F test of joint significance of all explanatory variables is $3.8 \cdot 10^{-8}$, which means that the null hypothesis of their joint insignificance is strongly rejected. As shown in Table 6.4, the variables PC_1 and PC_3 are both statistically insignificant at any reasonable significance level according to the t test. Thus an F test of the joint significance of all four PCs is carried out. The resulting p-value is 0.01 which confirms joint significance of the PCs. Therefore, all four PCs should be kept in the model as one would exclude unknown partial variance in original data when excluding PC_1 and PC_3 (Jolliffe, 2002).

Table 6.4: Care utilization per capita estimated by OLS

Variable	Coefficient estimate	P-value for t statistic
$intercept$	17.806131	0.000793
PC_1	0.003537	0.773268
PC_2	-0.041236	0.003719
PC_3	0.019172	0.410498
PC_4	0.051262	0.032385
$\log(PhysPC+1)$	0.318435	0.000021
$\log(AvAge)$	-3.998200	0.004727
$\log(AvNeighPhysPC)$	0.211100	0.011293

Source: Author's computations.

There is a high number of alternative specifications of this model with respect to the choice of explanatory variables and their functional form. Regression results of ten of these alternative specifications can be found in Tables B.1 through B.10 in Appendix B. In general, replacing $\log(AvAge)$ with another measure of age structure ($AvAge$, $\log(OAindex)$, $OAindex$) leads to lower fit based on R^2. Using the $AvAge$ does not worsen the fit and the models yield very similar results, however, $\log(AvAge)$ is kept for the sake of interpretation as a unit change in average age is a huge jump in value. The decision is much simpler for the case of $\log(AvNeighPhysPC)$ as there is always a drop in R^2 when it is replaced by either $AvNeighPhysPC$, $NeighPhysPC$, or $\log(NeighPhysPC)$. However, the results are robust when one of these measures is changed. If either $Pdens$ or $\log(Pdens)$ is added to the model, its coefficient is insignificant at any reasonable significance level. Moreover the fit measured by adjusted R^2 is worsened as the addition of either of these variables does not increase fit corrected for degrees of freedom. The same result is yielded when either $Unemp$ or $\log(Unemp)$ is added to the model.

The OLS model estimated in Table 6.4 needs to be tested for OLS assumptions, model specification, as well as for spatial effects. Model specification in terms of missing powers of explanatory variables is tested by the RESET test. The p-value of 0.66 shows that one cannot reject the null hypothesis of no missing powers of explanatory variables at any reasonable significance level. It can thus be concluded that no powers of explanatory variables are missing. Homoskedasticity of residuals was tested by the Breusch-Pagan test with p-value 0.79 and by the White's test with p-value 0.76. Thus, the null hypothesis of homoskedasticity of these tests cannot be rejected at any reasonable significance level and it can be concluded that the model satisfies the homoskedasticity assumption. Normality of residuals was tested by the Jarque-Bera test and the Shapiro-Wilk test resulting in p-values 0.26 and 0.28 respectively. Thus, one cannot reject the null hypotheses of the tests even at 20% significance level and one can conclude that the residuals from the model are normal.

Should spatial effects still take place in the model, the OLS method is, however, not appropriate even though the OLS assumptions hold. The OLS model, therefore, needs to be tested by spatial tests. The Moran's I test of residuals for spatial autocorrelation was performed on the OLS model. For the one-sided alternative hypothesis of positive spatial autocorrelation in residuals, the test yields a p-value of 0.02 and therefore one rejects the null of no spatial autocorrelation at 5% significance level in favor of positive spatial autocorrelation. If the two-sided test is used, the resulting p-value is 0.046 and no spatial autocorrelation is rejected in favor of the more general two-sided alternative of spatial autocorrelation at 5% significance level.

The result of the Moran's I test shows that there are residual spatial effects untreated in the OLS model. Therefore, the set of four LM tests is performed to assess what effects these might be. The LM test for an omitted spatially lagged dependent variable results in a p-value of 0.02, which means that the null hypothesis under which the lagged dependent variable is not omitted is rejected at 5% level. If the version of the test robust to omitted spatially autocorrelated residuals is performed, the null hypothesis is still rejected at 5% level as the test's p-value is 0.01. When using the LM test to test for spatially autocorrelated residuals, one obtains a p-value of 0.23 and fails to reject the null of no spatial autocorrelation in residuals. However, as a missing spatially lagged dependent variable was detected by the LM test, the robust version of the test for residuals should be used. The robust test results in a p-value of 0.08 and therefore one fails to reject the null hypothesis at 5% significance level and concludes that a structure of spatially autocorrelated residuals is not necessary to be included in the model. Hence the SAR model will be estimated and it will be tested for residual spatial autocorrelation as well because the null

hypothesis of the robust LM test would be rejected at 10% level and the evidence from this test is, therefore, not very strong.

The variables used for estimation of the SAR model were based on the explanatory variables of the OLS model. However, the variable $\log(AvNeighPhysPC)$ needs to be removed as it would describe a modified form of spatially lagged number of physicians per capita and is, therefore, not suitable as an explanatory variable in the SAR model. Nevertheless, a specification of SDM will be estimated as well in order to incorporate spatial lags of independent variables.

The SAR model estimated using variables based on OLS results is

$$\log(UtilPC) = \rho W \log(UtilPC) + \alpha\iota_n + \sum_{k=1}^{4} \beta_k PC_k + \beta_5 \log(PhysPC{+}1) + \beta_6 \log(AvAge) + \varepsilon$$

where the term $W \log(UtilPC)$ represents the spatially lagged dependent variable. The results of this estimation can be viewed in Table 6.5.

Table 6.5: Care utilization per capita estimated by SAR

Variable	Coefficient estimate	P-value of the Wald test
$intercept$	13.90865	0.00252
PC_1	-0.00285	0.79532
PC_2	- 0.02645	0.03307
PC_3	0.00095	0.96315
PC_4	0.05357	0.00893
$\log(PhysPC{+}1)$	0.26836	0.00002
$\log(AvAge)$	- 3.29910	0.00722
$W \log(UtilPC)$	0.44157	0.00004

Source: Author's computations.

All explanatory variables except for PC_1 and PC_3 are statistically significant at 5% according to the Wald test in the estimated SAR model in Table 6.5. As two of the PCs are insignificant for any reasonable significance level, it is necessary to test for the joint significance of all PCs. This is performed using the LR test against a nested model excluding the PCs as the test's null hypothesis. The p-value from the LR test is 0.03 which leads to rejection of the null hypothesis at 5% level and thus one can conclude that the PCs are jointly significant and should therefore be included in the model.

The significance of the ρ coefficient on the spatially lagged dependent variable was tested as well. The null hypothesis of $\rho = 0$ was tested against the alternative of $\rho \neq 0$ using the LR test, the LM test and the Wald test. The testing statistics together with corresponding p-values can be found in Table 6.6. The conclusion of all three tests is a strong rejection of the null hypothesis as all three p-values are well below the standard 5% significance level. One can thus conclude that the spatial lag of the dependent variable is strongly significant and that the SAR model is necessary to be estimated instead of the OLS model. Moreover, one can see in Table 6.6 that the theoretical inequality $Wald \geq LR \geq LM$ of testing statistics for the three tests holds for the model.

Table 6.6: Significance of the ρ coefficient

Test	Testing statistic value	P-value for the testing statistic
LR	12.898	0.00032889
LM	8.963	0.00275550
Wald	16.703	0.00004372

Source: Author's computations.

Residual spatial autocorrelation in the error term ε needs to be tested for as its presence would necessitate the estimation of a SAC model or a SARMA model instead of the SAR model. The LM test with its null hypothesis of no residual spatial autocorrelation is performed and the obtained p-value of the test is 0.47. Thus one cannot reject the null hypothesis of no residual spatial autocorrelation and concludes that the SAR model should be preferred to SAC or SARMA specifications.

The Nagelkerke pseudo-R^2 of the estimated SAR model is 0.53 and thus the fit of the model is satisfactory. The pseudo-R^2 value shows that 47% of variation in the dependent variable remains unexplained.

The maximized value of log-likelihood for the SAR model is 16.62465 and the corresponding value of AIC is -13.249. The AIC can serve for comparisons of the estimated SAR model with alternative model specifications. Several alternative SAR specifications were estimated and the estimation results are provided in Tables B.11 through B.17 in Appendix B. Replacing $\log(AvAge)$ by its alternatives $AvAge$, $OAindex$, and $\log(OAindex)$ leads to an increase in the AIC and a decrease in the maximized log-likelihood value and thus one can conclude that the model with $\log(AvAge)$ fits the data better. The addition of any of the variables $Pdens$, $\log(Pdens)$, $Unemp$, and $\log(Unemp)$ also leads to an increase in the AIC,

moreover, these variables are statistically insignificant at 5% level. Thus one concludes that none of these variables should be included in the SAR model.

Finally, the SAR model needs to be tested against the corresponding SDM model which includes spatial lags of explanatory variables. In order to do this, the SDM model needs to be estimated. The AIC of the SDM model is -10.235 which shows that the increase in log-likelihood is offset by the addition of six explanatory variables. If one compares the SAR and the SDM model by the LR test which tests the null hypothesis of the SAR against the alternative of SDM, one obtains a p-value of 0.16. Hence it is not possible to reject the null hypothesis even at 15% significance level and it can be concluded that adding the spatial lags of explanatory variables and using the SDM worsens the model as the increase in log-likelihood is offset by the decrease in degrees of freedom of the model. This is in accord with the conclusion obtained from the comparison of the AICs. In order to address the significance of $\log(AvNeighPhysPC)$ in the original OLS model, note that the spatial lag of $\log(PhysPC+1)$ is not significant in the SDM model (p-value from Wald test 0.12) and thus the significance of $\log(AvNeighPhysPC)$ in the OLS model was probably caused by the missing spatial lag of the dependent variable.

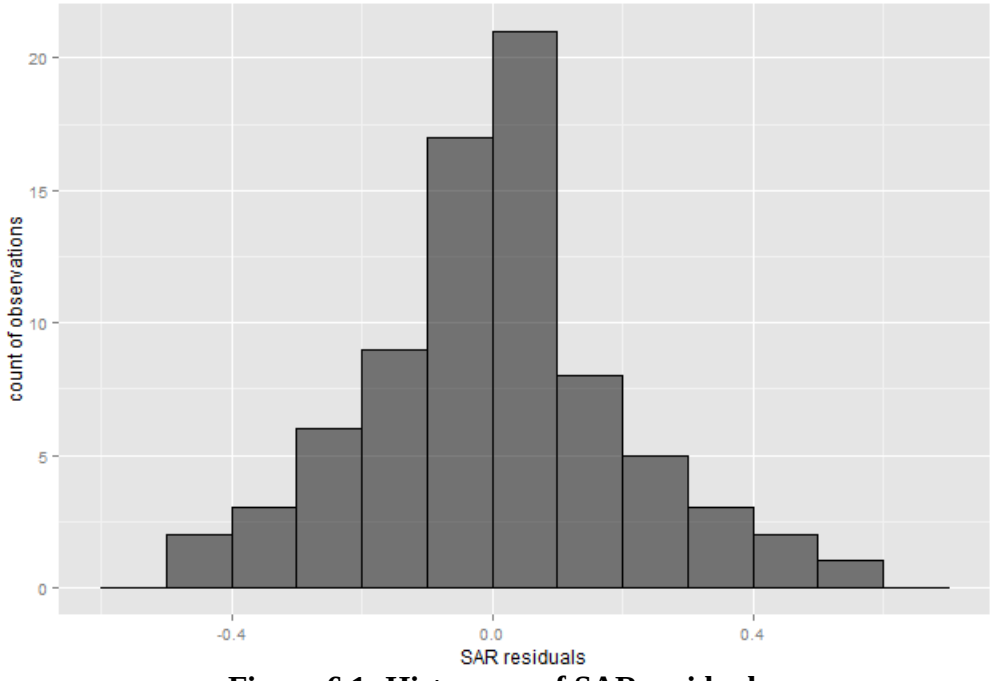

Figure 6.1: Histogram of SAR residuals

Source: Author's computations.

The SAR model needs to be tested for homoskedasticity and normality of residuals in order to make sure that the MLE estimation procedure can be applied. The p-value of the spatial studentized Breusch-Pagan test is 0.78 which implies that its null hypothesis of homoskedasticity cannot be rejected at any reasonable

significance level. It is thus concluded that the residuals are homoskedastic and no remedies are necessary. The normality of residuals is tested by the Jarque-Bera test and the Shapiro-Wilk test. Both tests fail to reject the null hypothesis of normality at any reasonable significance level as their p-values are 0.38 and 0.43 respectively. The assumption of normal distribution of the SAR residuals is thus concluded to be satisfied as well. The distribution of residuals from the SAR model is illustrated by their histogram in Figure 6.1.

6.2.4 Interpretation of Model Results

The coefficients of the SAR model cannot be interpreted in the usual way and therefore impacts must be calculated for each explanatory variable. Average direct impact, average indirect impact, and average total impact were computed for all explanatory variables of the SAR model. In order to be able to distinguish statistical significance of these impacts, their p-values were simulated using 1,000 iterations.

The estimated impacts together with their simulated p-values can be found in Table 6.7. The table contains the average direct impact, average indirect impact, and average total impact in columns with their p-values reported in parenthesis. The impacts will first be commented on quantitatively and then their qualitative meaning will be put into wider health-care context.

Table 6.7: Impacts from the SAR model of care utilization

Variable	Direct impact (simulated p-value)	Indirect impact (simulated p-value)	Total impact (simulated p-value)
PC_1	- 0.0030 (0.8209)	- 0.0021 (0.8130)	- 0.0051 (0.8135)
PC_2	- 0.0278 (0.0228)	- 0.01958 (0.1026)	- 0.0474 (0.0317)
PC_3	0.0010 (0.9387)	0.0007 (0.9578)	0.0017 (0.9463)
PC_4	0.0563 (0.0090)	0.0397 (0.0648)	0.0959 (0.0125)
$\log(PhysPC+1)$	0.2819 (0.0000)	0.1986 (0.0422)	0.4806 (0.0008)
$\log(AvAge)$	- 3.4658 (0.0041)	- 2.4420 (0.0741)	- 5.9078 (0.0108)

Note: Simulation performed using 1,000 iterations.
Source: Author's computations.

Next, the model interpretation will be discussed, starting with the impacts of $\log(PhysPC+1)$. As the impacts are interpreted in the usual partial-derivative manner, one would like to interpret them as the influence of a percentage change in $PhysPC+1$ on a percentage change in $UtilPC$. However, although the use of $\log(PhysPC+1)$ was helpful for estimation purposes, it rather complicates the model interpretation as a percentage change in $PhysPC+1$ is not equivalent to a percentage change in $PhysPC$. Nevertheless, Wooldridge (2009, p. 192) states that: "Generally, using $\log(1+y)$ and then interpreting the estimates as if the variable were $\log(y)$ is acceptable when the data on y contain relatively few zeros." As the data on $PhysPC$ contain only 2 zeros out of 77 observations, the interpretation of $\log(PhysPC+1)$ as $\log(PhysPC)$ can be used although one must keep in mind that this interpretation is not exact. This, however, is not a fatal problem as exact interpretation of coefficients is not the purpose of this work. The important fact is that the qualitative results are unchanged.

The interpretation of the impact of $\log(PhysPC+1)$ therefore is that if the number of ACI physicians per 100,000 inhabitants increases by 1% then the per capita number of utilized points increases ceteris paribus by 0.48%. This effect is strongly significant. The direct effect of this change is a 0.28% increase in points utilization per capita and it is also strongly significant. The indirect effect resulting from feedback loops is 0.20% and it is also significant at 5% level.

The interpretation of the impacts of $\log(AvAge)$ is quite straightforward as the percentage change interpretation can be applied. The total effect of a 1% increase in the average age is associated with a 5.91% ceteris paribus decrease in the per capita number of utilized points. The total effect is strongly significant. When divided between the direct and the indirect effect, the direct effect is a 3.47% decrease and it is also strongly significant. On the other hand, the indirect effect of this change is a 2.44% decrease and it is statistically insignificant at 5% level.

The quantitative impact of PCs can be computed using the exact general log-level interpretation formula by Wooldridge (2009)

$$\widehat{\%\Delta y} = 100 \cdot \left[\exp\left(\hat{\beta} \cdot \Delta x\right) - 1\right]$$

which describes the effect of a level change in an explanatory variable x with an estimated coefficient $\hat{\beta}$ as a percentage change in the dependent variable y. When applied to the direct impacts of PCs, the resulting ceteris paribus changes in the per capita points utilization can be computed and these are provided in Table 6.8.

As the p-values for total impacts of PC_1 and PC_3 are very close to one, the interpretation of these variables is practically redundant. In the case of PC_2 one can see in Table 6.8 that its unit increase leads to a statistically significant 4.63% decrease in care utilization. On the other hand, a unit increase in PC_4 leads to a 10.07% increase in care utilization in points per capita. However, it is crucial to realize that although these effects seem huge compared to the effects of changes in age or number of physicians, one must keep in mind that a unit change in a PC is an enormous one unlikely ever to take place. Thus, the economic significance of these effects should not be considered of greater importance than the one of other variables. As the economic interpretation of PCs is in general complicated, it seems redundant to provide an in-depth analysis of their direct and indirect effects.

Table 6.8: Total impact of environmental PCs

Principal component	Total impact estimate	The effect of a unit increase on care utilization	Statistical significance
PC_1	-0.0051	- 0.51%	Practically none.
PC_2	-0.0474	- 4.63%	Significant at 5%.
PC_3	0.0017	+ 0.17%	Practically none.
PC_4	0.0959	+ 10.07%	Significant at 5%.

Source: Author's computations.

Unlike in the case of explanatory variables, the coefficient on the spatial lag of the dependent variable can be interpreted directly without the use of impacts. The estimated value of the coefficient (0.44157) was provided in Table 6.5 and its strong statistical significance was assessed in Section 6.2.3. Based on the Wooldridge (2009) interpretation formula, the quantitative interpretation of this coefficient is that if the average of logged points for care utilization per capita in contiguous districts increases by one, the number of points for care utilization per capita increases by 55.51%. Once again, the probability of such large change in the average of $\log(UtilPC)$ in contiguous districts is very unlikely as $\log(UtilPC)$ ranges from 2.93 to 4.09 in the sample. Hence the economic importance of the spatial lag coefficient should not be exaggerated. If one considers a much more likely 0.1 change in the averaged $\log(UtilPC)$ over contiguous districts, the estimated effect is a 4.51% increase only.

The next point of this section is the much more remarkable economic interpretation of the impacts. The positive direct impact of number of physicians is expected as when the number of physicians increases in a district, then care availability increases and thus one expecst care utilization to increase as well. This

might be due to the fact that there were too few ACI physicians in the first place which led to underuse of care. Furthermore, it might also be due to the fact that later there are too many of them which could potentially lead even to overproduction.

The direct impact of average age is negative, which might be a contra intuitive result for health care in general, however, as a large proportion of ACI patients are children, it is not surprising. Districts with lower average age are likely to have a higher share of children in the population and therefore these districts exhibit higher care utilization.

The interpretation of the direct impact of principal components is rather complicated. PC_2 could be interpreted as follows: if there is more natural plant life (forests, grassland, protected areas) and less artificial plant life (arable land), then care utilization is lower. This might show that the pollen produced by plants is not as big of a problem as the high amounts of pesticides used nowadays on the majority of arable land. It is argued that many pesticides are toxic and that they cause allergic and also other diseases (Medical University of South Carolina, 2015). The impact of PC_4 could be interpreted as an increase in care utilization as a result of more roads and smaller or fewer water body areas and hop gardens and vineyards. It is expected that roads increase care utilization as vehicles pollute environment and it is also expected that water body areas decrease care utilization as these do not produce any allergens. It is necessary, however, to keep in mind that this interpretation is neither precise nor an important output of the work.

The indirect effects of explanatory variables are of the same directions as the direct effects. This is caused by the positive spatial lag coefficient ρ in the model. If e.g. utilization in one district increases due to a change in an explanatory variable, utilization in neighboring districts will also increase. This way, the feedback loop effect will be transferred through all neighboring relations back to the original district where an explanatory variable changed. The effects among neighbors will be positive and thus the indirect effect will be of the same sign as the direct effect.

Interpreting the spatial lag coefficient itself is often a challenge in spatial econometrics. The SAR model estimates that if utilization of care increases in one district, it increases also in neighboring districts. This might be caused by disease awareness, which can be shown using a simple example: a child in district A eats peanuts for the first time in her life, the child has a strong allergic reaction to it and is transported to a hospital, diagnosed with peanut allergy and from then on utilizes ACI care. The child's mother tells this story to all her friends some of whom live in neighboring districts. They all want to prevent this happening to their own children

and visit ACI physicians with them in order to find out about their allergies. Of course, this is a very simplified individual case, however, it can show us that information about treated ACI diseases can spread among people (and naturally not only within a district) and many others might decide to learn about their health status in more detail. And the more people utilize care, the more likely information is to spread among others.

The causal relationship in care utilization, nevertheless, does not necessarily have to be a direct one in the sense that utilization itself would induce further utilization in neighboring districts. This relationship might also be caused by the fact that utilization levels are a factor people can (indirectly) observe. Consider two districts A and B that are contiguous to one another and consider that there is a lack of ACI physicians in district B and thus some patients from district B actually utilize care by visiting a physician in district A (this care will be counted towards district B as we are dealing with the dataset for care utilization). Now consider the situation that the number of physicians in district B increases. Besides the fact that care becomes more accessible for patients from district B and that overall utilization in district B will probably increase, this might also cause that fewer patients from district B will now utilize care by visiting physicians in district A. As a result of lower workload of physicians in district A, there is an opportunity for patients from district A to utilize a higher amount of care as e.g. waiting times for a visit are likely to shorten. In this case, care utilization in district A seems to be caused by a higher utilization in district B because the majority of patients do not observe the number of physicians in districts directly. What patients do observe, however, is e.g. the waiting time for an appointment which can be influenced by care utilization levels in neighboring districts.

However, it might also be the case that the positive relationship of care utilization in neighboring districts is not causal. Bivand et al. (2008, p. 273) explain that "it is quite often the case that observations on relevant covariates are not available at all, and that the detection of spatial autocorrelation in data or model residuals in fact constitutes the only way left to model the remaining variation". Thus the positive coefficient might as well show that there are some unobserved variables which are spatially autocorrelated and influence care utilization. The effect of these unobserved variables is captured by the spatial lag coefficient in the SAR model. Thus, the spatial model allows one to partially explain variation in care utilization by utilization in neighboring districts. If a spatial model was not implemented, this variation would remain unexplained. And although in this case it is not possible to

draw any causal conclusions from the spatial lag coefficient, it still increases the model's explanatory power and is thus useful in this work.

The resulting spatial lag coefficient could incorporate one or all of the three above-mentioned effects. This might be a motivation for institutions to collect more detailed data on ACI in the future and enable researches to draw more certain conclusions about practical meaning of the coefficient. However, as a wide variety of explanatory variables was controlled or tested for, the space for possible latent factors is quite limited.

Finally, the fitted values of care utilization from the SAR model will be explored. As the fitted values from the model are $\log(\widehat{UtilPC})$ and our interest lies in $\widehat{UtilPC}$, it is necessary to recalculate the log-fitted values into levels. Wooldridge (2009) claims that this is not possible to be done only by exponentiation of $\log(\widehat{UtilPC})$ as the result would be biased. Based on the general procedure by Wooldridge (2009), the fitted values for care utilization can be obtained as

$$\widehat{UtilPC} = \exp\left(\log(\widehat{UtilPC})\right) \cdot n^{-1} \cdot \sum_{i=1}^{n} \exp(\hat{\varepsilon}_i)$$

where $n = 77$ is the number of observations in the model and $\hat{\varepsilon}$ are the SAR model residuals. Hence the fitted values of care utilization are calculated and can be compared to the observed values.

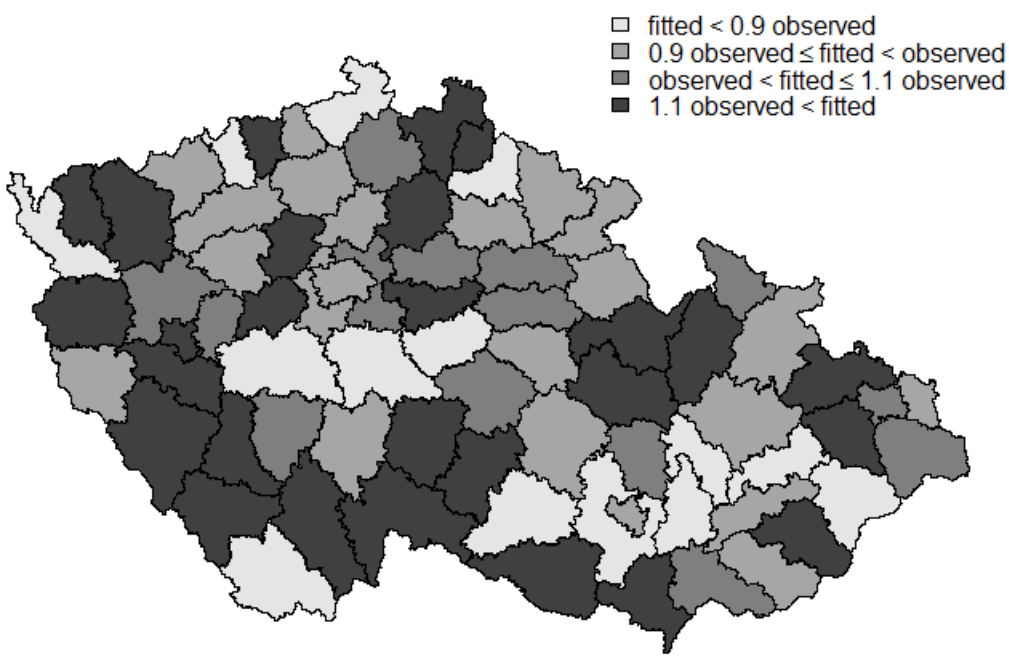

Figure 6.2: Care utilization - comparison of observed and fitted values

Source: Author's computations, shapefile by URRlab (2013).

Figure 6.2 compares the fitted $\widehat{UtilPC}$ to the observed $UtilPC$. The districts on the map are divided into four groups. The first group (in lightest gray) shows districts where the fitted value is below 90% of the observed value. These are the district for which the model predicts distinctly lower levels of care utilization than the real 2012 values. Thus, this could show areas where savings on ACI care might be possible. The 13 districts in which the fitted value is distinctly lower than the observed one are: Cheb, Most, Děčín, Semily, Příbram, Benešov, Kutná Hora, Třebíč, Brno-country, Vyškov, Prostějov, Přerov, and Vsetín.

The other extreme is the group of districts (in darkest gray) where the fitted value is above 110% of the observed value. These are the districts where there is a possibility of underuse of care. The 27 districts in this group are: Sokolov, Karlovy Vary, Teplice, Liberec, Jablonec nad Nisou, Mladá Boleslav, Kladno, Beroun, Kolín, Tachov, Plzeň-town, Plzeň-south, Klatovy, Strakonice, Prachatice, České Budějovice, Jindřichův Hradec, Pelhřimov, Jihlava, Znojmo, Břeclav, Zlín, Opava, Nový Jičín, Ústí nad Orlicí, Svitavy, and Šumperk.

The remaining 37 districts have fitted values between 90% and 110% of observed values. Thus these are the districts where the model predicts values quite similar to the observed ones.

The outcomes of the utilization model bring up several intriguing issues. First, according to the SAR model, 47% of variation in care utilization still remains unexplained after controlling for all available theoretical determinants of care utilization. Thus, one must consider the possibility that some of the variation in care utilization is unwarranted and might thus show unresolved inequity or inefficiency in the entire system.

Second, the strongly significant relationship between care utilization and the number of physicians could be a result of either insufficient number of ACI physicians in the system or their non-optimal geographical distribution. Česká společnost alergologie a klinické imunologie (n.d.) recommends that there should be four ACI physicians per 100,000 inhabitants and although in the CR, there are 4.17 of them (ÚZIS ČR, 2013; Czech Statistical Office, 2015a), the median value over districts is only 2.80 ACI physicians per 100,000 inhabitants. Therefore, although the overall number of ACI physicians is sufficient, their geographic distribution causes their very low number in some districts (even zero in two of them). As in the majority (80.52%) of districts the number of ACI physicians is below the four-ACI-physicians benchmark, in general, there might be shortage of ACI physicians causing the significance as physicians in those districts work on top of their capacities and their

limited number prevents patients from sufficient care utilization. This is supported by the fact that although 30-40% of Czech population suffer from ACI diseases, only 8.5% of Czech population utilized ACI care in 2012 as already shown in Section 2.1.

Third, a simultaneous relationship of care utilization in neighboring districts was found, although this is not necessarily a causal one as described before. This relationship, however, indicates that if the policymakers wish to influence care utilization, a change in utilization in one district is likely to induce changes in utilization in other districts as well. This is, however, double-edged. This relationship might work in favor of the policymaker e.g. when the policymaker decides to lower care utilization in multiple districts. It might then be sufficient to imply changes in only a few of them and utilization in the rest of them would decrease as well at no costs. On the other hand, if the policy is not unified, it might happen that differing policies are implemented in neighboring districts and their effects might partially cancel out thus wasting resources. The SAR model therefore shows that it is favorable to coordinate health policy in the branch of ACI. Coordination, however, is implemented only partially as currently it is the regional authorities who register providers (as described in Section 2.3). In the light of the SAR model, this situation is clearly more favorable than no coordination at all. However if one considers the map of Czech regions and districts (Figure A.1, Appendix A), one can see that a high number of districts is contiguous to district(s) belonging to another region. Thus the situation that districts with different policies affect each other due to their contiguity is quite likely. Based on the SAR model a countrywide coordination would be preferable.

6.3 Care Provision

The estimation of care provision is based on the idea that care provision is related to the physician who provides care. Thus the dependent variable in the model should be points provided in ACI per ACI physician ($ProvPP$). However, this variable is not defined for 2 of the 77 districts because of zero number ACI of physicians. Therefore, these observations cannot be used for estimation and will be dropped from all explanatory variables including the columns and rows of the contiguity matrix W. The model for care provision will, therefore, be estimated using 75 observations.

As the variable $ProvPP$ is not normally distributed (Jarque-Bera test p-value below $2.2 \cdot 10^{-16}$, Shapiro-Wilk test p-value $3.5 \cdot 10^{-10}$) and is positively skewed (skewness parameter 3.23) the recommendation by Wooldridge (2009) is again applied and its natural logarithm is computed. Nevertheless, the new variable $\log(ProvPP)$ is

also not normally distributed although the p-values increase for both tests ($2.3 \cdot 10^{-6}$ for the Jarque-Bera test, 0.01 for the Shapiro-Wilk test). Hence $\log(ProvPP)$ seems more suitable for estimation, however, models including both the level and the log will be estimated.

The choice of explanatory variables is based on several principles. First, variables that might describe need for care by patients are included. These are the environmental PCs (PC_1, PC_2, PC_3, PC_4) and measures of age ($AvAge, OAindex$). They are included because it might be the case that if need for care is large, physicians might provide more care under pressure of patients. Second, variables describing the population density ($Pdens$) and unemployment rate ($Unemp$) are included as these might approximate the attractiveness of the district based on which the physician might make the choice where to live and work. This might then cause undersaturation or oversaturation of districts by ACI physicians and have an impact on the care they are able to provide. Third, the percentage of points provided by ACI physicians ($pctgACI$) is be included in the model in order to incorporate the possible effect that ACI physicians in some districts might be overloaded and that patients are, therefore, forced to obtain ACI care from other types of physicians who provide it instead.

The discussion on functional forms performed in Section 6.2.3 in the utilization model will not be repeated here. Dropping two observations from the dataset does not influence the use of natural logarithms. The additional variable $pctgACI$ is not normally distributed (p-values $1.3 \cdot 10^{-9}$ for the Shapiro-Wilk test, below $2.2 \cdot 10^{-16}$ for the Jarque-Bera test), however, its log-transformation $\log(pctgACI)$ does not help as the p-value of the Shapiro-Wilk test actually drops to $1.6 \cdot 10^{-13}$. Note that the observation for which $pctgACI$ is not defined was dropped from the dataset due to $ProvPP$.

In the first step, 22 OLS models for both forms of the dependent variable were estimated. Each model has the intercept and one explanatory variable. The models were estimated both for the log and level forms of all explanatory variables. The only exceptions are the two models using PCs which each include the intercept and the four PCs in level form.

The results of all 22 OLS models are provided in Table 6.9. The used dependent and independent variables are set in the first two columns for each model. The estimated regression coefficient on the independent variable(s) is reported in the third column. In the fourth column, the p-value for the t test of significance is provided for each explanatory variable. An F test of joint significance was performed for each model and its p-value is reported in column five. For the models with only

one explanatory variable, this yields the same result as the t test, however, it is useful in the case of models including the PCs. Both tests were included for the sake of completeness although in the majority of cases, they are statistically equivalent. Columns six and seven then report the regression R^2 and adjusted R^2. Please note that the two measures are not comparable for models with different dependent variables (Wooldridge, 2009).

Table 6.9: Care provision - OLS estimation of multiple specifications

Dependent variable	Independent variable(s)	Estimated coefficient	P-value of t test	P-value of F test	R^2	Adjusted R^2
$ProvPP$	$AvAge$	- 25970	0.769	0.769	0.001	- 0.012
$ProvPP$	$\log(AvAge)$	- 1091284	0.761	0.761	0.001	- 0.012
$ProvPP$	$OAindex$	- 2228	0.683	0.683	0.002	- 0.011
$ProvPP$	$\log(OAindex)$	- 251085	0.658	0.658	0.003	- 0.011
$ProvPP$	$Unemp$	- 5194	0.882	0.882	0.000	- 0.013
$ProvPP$	$\log(Unemp)$	- 59622	0.801	0.801	0.001	- 0.013
$ProvPP$	$Pdens$	- 27577	0.154	0.154	0.028	0.014
$ProvPP$	$\log(Pdens)$	- 154225	0.084	0.084	0.040	0.027
$ProvPP$	$pctgACI$	- 179825	0.707	0.707	0.002	- 0.012
$ProvPP$	$\log(pctgACI)$	- 55964	0.841	0.841	0.001	- 0.013
$ProvPP$	PC_1	- 26094	0.405	0.216	0.004	0.026
	PC_2	- 37769	0.314			
	PC_3	- 91222	0.109			
	PC_4	72235	0.214			
$\log(ProvPP)$	$AvAge$	- 0.026	0.585	0.585	0.004	- 0.010
$\log(ProvPP)$	$\log(AvAge)$	-1.096	0.577	0.577	0.004	- 0.009
$\log(ProvPP)$	$OAindex$	- 0.002	0.455	0.455	0.008	- 0.006
$\log(ProvPP)$	$\log(OAindex)$	- 0.243	0.432	0.432	0.008	- 0.005
$\log(ProvPP)$	$Unemp$	- 0.005	0.805	0.805	0.001	- 0.013
$\log(ProvPP)$	$\log(Unemp)$	- 0.048	0.714	0.714	0.002	- 0.012
$\log(ProvPP)$	$Pdens$	- 0.018	0.091	0.091	0.039	0.025
$\log(ProvPP)$	$\log(Pdens)$	- 0.094	0.053	0.053	0.051	0.038
$\log(ProvPP)$	$pctgACI$	- 0.188	0.474	0.474	0.007	- 0.007
$\log(ProvPP)$	$\log(pctgACI)$	- 0.076	0.619	0.619	0.003	- 0.010
$\log(ProvPP)$	PC_1	- 0.017	0.315	0.113	0.010	0.048
	PC_2	- 0.014	0.492			
	PC_3	- 0.056	0.072			
	PC_4	0.053	0.093			

Source: Author's computations.

When comparing the regression results for the models with $ProvPP$ as the dependent variable, it may be observed that the vast majority of models is practically useless. The R^2 is well below 1%, the adjusted R^2 is negative and the t test and F test report p-values close to one, which means that the explanatory variables are not significant at any reasonable significance level. There are three exceptions from this rule. The model using $Pdens$ and the model using the PCs as explanatory variables, however still driven (jointly) insignificant by the F test, at least have a positive adjusted R^2. The only model with explanatory variable significant at 10% is the one using $\log(Pdens)$, which explains about 4% of the variation in $ProvPP$.

When focusing on the models using $\log(ProvPP)$ as the dependent variable, the general results are very similar. Although the significance increases slightly as the F test p-values are further from one, still the majority of used explanatory variables is insignificant, R^2 is below 1% and the adjusted R^2 is negative. The model using PCs as explanatory variables has positive adjusted R^2 and explains about 1% of the variation in the dependent variable. However, the PCs are jointly insignificant at 10%. The model with $Pdens$ shows significance of the explanatory variable at 10% and explains about 4% of variation in $\log(ProvPP)$. The only model whose explanatory variable is significant close to 5% level is the one using $\log(Pdens)$, which explains about 5% of variation in $\log(ProvPP)$.

Although it does not make much sense to dwell in such poorly-explaining models, those with positive adjusted R^2 were tested for remaining spatial autocorrelation in the error term as well as for the omitted spatially lagged dependent variables using the series of four LM tests. Not surprisingly, as the variable $ProvPP$ itself is not spatially autocorrelated, these tests resulted in no rejection of the hypotheses that such effects are missing in the models.

Thus, it can be concluded that care provision per physician cannot be modeled using the data that are currently available. An interesting outcome of this section, however, is that variables that might influence the need for care of patients do not influence care provision at all. The amount of care provided by ACI physicians is clearly determined by other factors. This is a valid motivation to conduct a Medical Practice Variation study in the branch of ACI in the future and investigate whether physicians' believes or patients' needs are prevailing when care provided is determined. For such study, however, a much more detailed and preferably individual-level dataset would be necessary and the author hopes that its collection will be possible.

7 Conclusion

The branch of ACI is a very important one as diseases related to it involve a large part of the Czech population. This thesis, therefore, explored the care utilization and provision in the context of geographic variation while applying methods of spatial econometrics. Thus, a gap in the existing ACI literature is aimed to be filled, as no previous works on this topic exist.

A detailed analysis of the data on care utilization and provision was conducted with the goal of their comparison as this is the first use of the dataset. It was found that although both utilization and provision vary greatly among districts, there are several differences between the two datasets. One of them is that some of the ACI health services are provided only in a very limited number of districts while utilized by patients from a much larger number of districts. Essentially, this means that patients travel to utilize at least some of the ACI health services.

Spatial autocorrelation of several variables was tested because the presence of spatial autocorrelation shows that neighboring districts exhibit similar levels of a variable, while its absence shows that neighboring districts have different values. It was found that care utilization per capita is strongly spatially autocorrelated while care provision per capita is not. Moreover, it was uncovered that the number of ACI physicians per capita is not spatially autocorrelated either. On the contrary, physicians tend to concentrate in bigger cities while there are districts where there is no ACI physician. This finding shows that the distribution of ACI physicians and the care that they provide does not geographically correspond to the utilization needs of patients.

Care utilization per capita was modeled using a spatial autoregressive model specification. This model, while controlling for the per capita number of physicians, the age structure, and measures of environment, explained about half of the variation in per capita care utilization. Fitted values from the model were then compared to observed values of per capita care utilization. Out of the total of 77 Czech districts, the per capita care utilization was distinctly higher than the one predicted by the model in 13 of them, while it was distinctly lower in 27 districts. These are the districts that are thus most likely to exhibit inefficiencies or inequity in ACI care utilization. A very significant relationship between per capita care utilization and per capita number of ACI physicians was found in the model. This, coupled with the fact

that the per capita number of physicians is lower than recommended in more than 80% of districts, shows that the non-optimal geographical distribution of ACI physicians could cause a serious shortage of capacities at which people can utilize care in several districts and that care utilization might simply be limited by this shortage. This is a serious problem if the fact that approximately a third of Czech population suffers from ACI diseases is taken into account.

The spatial autoregressive model also shows that even after controlling for several explanatory variables, the values of per capita care utilization are still spatially autocorrelated among districts. Although it is not possible to assess whether or not this is a causal relationship using the data that are available, this relationship might show the necessity of coordination in health policy. If there is indeed causality in this relationship, policy in neighboring districts should be coordinated, otherwise effects of differing policies might partially cancel each other out. To confirm this effect, however, further research needs to be conducted.

Attempts to estimate a model describing care provision per ACI physician were made. However, none of the variables that were available to the author were found to explain care provision per ACI physician. Interestingly not even variables like environment or age structure that are likely to influence the need for care by patients had any effect. The conclusion of the attempts to model care provision per ACI physician is that provision is probably determined by other factors for which data are not available. This shows the necessity of additional data collection in this area in order to be able to conduct future research. Moreover, the lack of impact of variables influencing the needs of patients posts several future research questions about the determinants of care provision.

Despite having disclosed the inappropriateness of geographical distribution of ACI physicians for care utilization, the present thesis should be followed by other works that would study geographic variation in Czech ACI in a more profound way. However, additional data need to be collected, especially for studies of care provision.

Bibliography

Akdis, C.A. & Agache, I. (eds.). (2014): *GLOBAL ATLAS OF ALLEGRY*. [Online] Zurich: European Academy of Allergy and Clinical Immunology. Available from: http://www.eaaci.org/resources/scientific-output/global-atlas-of-allergy.html [Accessed 2015-05-05].

Alexa, J., Rečka, L., Votápková, J., van Ginneken, E., Spranger, A. & Wittenbecher, F. (2015): Czech Republic: Health system review. *Health Systems in Transition*. [Online] vol. 17, no. 1, p. 1-165. Available from: http://www.euro.who.int/en/about-us/partners/observatory/publications/health-system-reviews-hits/countries-and-subregions/czech-republic-hit-2015 [Accessed 2015-06-22]. ISSN 1817-6127.

Anselin, L. (1988): *Spatial econometrics: methods and models*. Studies in operational regional science. Dordrecht: Kluwer. ISBN 90-247-3735-4.

Bartra, J., Mullol, J., Del Cuvillo, A., Dávilla, I., Ferrer, M., Jáuregui, I., Montoro, J., Sastre, J. & Valero, A. (2007): Air Pollution and Allergens. *Journal of Investigational Allergology and Clinical Immunology*. [Online] vol. 17, p. 3-8. Available from: http://www.jiaci.org/issues/vol17s2/2.pdf [Accessed: 2015-01-30].

Bivand, R. S., Pebesma, E. J. & Gómez-Rubio, V. (2008): *Applied Spatial Data Analysis with R*. Use R!. New York: Springer. ISBN 978-0-387-78170-9.

Bousquet, P.-J., Chinn, S., Janson, C., Kogevinas, M., Burney, P. & Jarvis, D. (2007): Geographical variation in the prevalence of positive skin tests to environmental aeroallergens in the European Community Respiratory Health Survey I. *Allergy*. [Online] vol. 62, issue 3, p. 301-309. Available from: http://onlinelibrary.wiley.com/doi/10.1111/j.1398-9995.2006.01293.x/full [Accessed 2015-12-04]. DOI 10.1111/j.1398-9995.2006.01293.x.

Bruni, M. L. & Mammi, I. (2015): *Spatial effects in hospital expenditures: a district level analysis*. Universita' di Bologna Working Papers. [Online] p. 24. Available from: https://ideas.repec.org/p/bol/bodewp/wp1027.html [Accessed 2015-12-04]. ISSN 2282-6483.

Česká společnost alergologie a klinické imunologie. (n.d.): *Koncepce oboru alergologie a klinická imunologie.* [Online]. Available from: http://www.csaki.cz/koncepce [Accessed 2015-06-22].

Český statistický úřad. (2014): *Číselník okresů (OKRES_LAU)*. [Online]. Available from: https://www.czso.cz/csu/czso/ciselnik_okresu_lau1_nuts_2008 [Accessed 2015-06-16].

Český statistický úřad. (2015): *Veřejná databáze: Průměrný podíl nezaměstnaných osob podle okresů ČR.* [Online]. Available from: http://vdb.czso.cz/vdbvo/tabdetail.jsp?kapitola_id=924&potvrd=Zobrazit+tabulku&go_zobraz=1&cislotab=PRA6010PC_OK&childsel0=1&voa=tabulka&str=tabdetail.jsp [Accessed 2015-06-17].

ČPZP. (2009): *Alergii často způsobuje oblíbené jídlo.* [online]. Available from: http://www.cpzp.cz/clanek/347-0-Alergii-casto-zpusobuje-oblibene-jidlo.html [Accessed: 2015-01-30].

Czech Republic. (2011a): Předpis č. 411/2011 Sb.: Vyhláška, kterou se mění vyhláška Ministerstva zdravotnictví č. 134/1998 Sb., kterou se vydává seznam zdravotních výkonů s bodovými hodnotami, ve znění pozdějších předpisů. *Sbírka zákonů.*

Czech Republic. (2011b): Předpis č. 425/2011 Sb.: Vyhláška o stanovení hodnot bodu, výše úhrad zdravotní péče hrazené z veřejného zdravotního pojištění a regulačních omezení objemu poskytnuté zdravotní péče hrazené z veřejného zdravotního pojištění pro rok 2012. *Sbírka zákonů.* [Online] p. 5506-5584. Available from: http://aplikace.mvcr.cz/sbirka-zakonu/SearchResult.aspx?q=425/2011&typeLaw=zakon&what=Cislo_zakona_smlouvy [Accessed 2015-06-22].

Czech Statistical Office. (2012): *Small Lexicon of Municipalities of the Czech Republic - 2012*. [Online]. Available from: https://www.czso.cz/csu/czso/small-lexicon-of-municipalities-of-the-czech-republic-2012-bo3txm7188 [Accessed 2015-08-27].

Czech Statistical Office. (2015a): *Public Database: Population in districts since 1991*. [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabparam.jsp?cislotab=DEMPU001_OK&kapitola_id=368&voa=tabulka&go_zobraz=1&aktualizuj=Aktualizovat&cas_1_88=20121231 [Accessed 2015-05-04].

Czech Statistical Office. (2015b): *Public Database: Population - average age, index of ageing.* [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabdetail.jsp?potvrd=Show+table&go_zobraz=1&cislotab=OBY5082PU_OK&cas_1_100=20121231&vo=null&voa=tabulka&str=tabdetail.jsp [Accessed 2015-06-17].

Czech Statistical Office. (2015c): *Public Database: Area - structure of land types.* [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabdetail.jsp?potvrd=Zobrazit+tabulku&cas_2_108=20121231&go_zobraz=1&cislotab=RSO5012PU_OK&vo=null&voa=tabulka&str=tabdetail.jsp [Accessed 2015-06-17].

Czech Statistical Office. (2015d): *Public Database: Emissions of main pollutants (REZZO 1-3).* [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabdetail.jsp?potvrd=Zobrazit+tabulku&go_zobraz=1&cislotab=ZPR5012PU_OK&vo=null&voa=tabulka&str=tabdetail.jsp&cas_1_103=2012 [Accessed 2015-06-18].

Czech Statistical Office. (2015e): *Public Database: Protected areas: area.* [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabdetail.jsp?kapitola_id=10&potvrd=Zobrazit+tabulku&go_zobraz=1&cislotab=ZPR5042PU_OK&cas_1_105=20121231&vo=null&voa=tabulka&str=tabdetail.jsp [Accessed 2015-06-18].

Czech Statistical Office. (2015f): *Public Database: Road Transport Infrastructure.* [Online]. Available from: http://vdb.czso.cz/vdbvo/en/tabdetail.jsp?potvrd=Zobrazit+tabulku&go_zobraz=1&cislotab=DOP5012PU_OK&vo=null&cas_1_101=20121231&voa=tabulka&str=tabdetail.jsp [Accessed 2015-06-18].

de Monchy, J. G., Demoly, P., Akdis, C. A., Cardona, V., Papadopoulos, N. G., Schmid-Grendelmeier, P. & Gayraud, J. (2013): Allergology in Europe, the blueprint. *Allergy.* [Online], vol. 68, issue 10, p. 1211-1218. Available from: http://onlinelibrary.wiley.com/doi/10.1111/all.12225/abstract [Accessed 2015-01-22]. DOI 10.1111/all.12225.

Eibich, P. & Ziebarth, N. R. (2014a): Analyzing regional variation in health care utilization using (rich) household microdata. *Health Policy.* [Online], vol. 144, issue 1, p. 41-53. Available from: http://www.sciencedirect.com/science/article/pii/S0168851013001127 [Accessed 2015-12-04]. DOI 10.1016/j.healthpol.2013.04.015.

Eibich, P. & Ziebarth, N. R. (2014b): Examining the structure of spatial health effects in Germany using Hierarchical Bayes Models. *Regional Science and Urban Economics.* [Online], vol. 49, p. 305-320. Available From: http://www.scienced

irect.com/science/article/pii/S016604621400060X [Accessed 2015-12-04]. DOI 10.1016/j.regsciurbeco.2014.06.005.

European Commission. (2015): *The 2015 Ageing Report: Economic and budgetary projections for the 28 EU Member States (2013-2060).* European Economy 3/2015. [Online] Luxembourg: Publications Office of the European Union. Available from: http://ec.europa.eu/economy_finance/publications/european_economy/2015/ee3_en.htm [Accessed 2015-06-17]. ISSN 1725-3217.

Filippini, M., Masiero, G. & Moschetti, K. (2009): Small area variations and welfare loss in the use of outpatient antibiotics. *Health Economics, Policy and Law.* [Online] vol. 4, issue 1, p. 55-77. Available from: https://doc.rero.ch/record/11238/files/filippini_HEPL_2009.pdf [Accessed 2015-12-04]. DOI 10.1017/S174413310800460X.

Greene, W. H. (2012): *Econometric Analysis.* 7th ed. Boston: Prentice Hall. ISBN 0-13-139538-6.

Gupta, R. S., Zhang, X., Sharp, L. K., Shannon, J. J. & Weiss, K. B. (2008): Geographic variability in childhood asthma prevalence in Chicago. *Journal of Allergy and Clinical Immunology.* [Online], vol. 121, issue 3, p. 639-645. Available from: http://www.sciencedirect.com/science/article/pii/S0091674907035725 [Accessed 2015-12-04]. DOI 10.1016/j.jaci.2007.11.036.

Jolliffe, I. T. (2002): *Principal Component Analysis, Second Edition.* Springer Series in Statistics. [Online] New York: Springer-Verlag, Inc. ISBN 0-387-95442-2. Available from: http://site.ebrary.com.ezproxy.is.cuni.cz/lib/cuni/detail.action?docID=10047693 [Accessed 2015-11-03].

Kašková, R. (2013a): *Allergology and Clinical Immunology - Activity in Branch in the Královéhradecký Region in 2012.*[Online]. Available from: http://www.uzis.cz/en/fast-information/allergology-and-clinical-immunology-activity-branch-kralovehradecky-region-2012 [Accessed 2015-06-16].

Kašková, R. (2013b): *Allergology and Clinical Immunology - Activity in Branch in the Liberecký Region in 2012.* [Online]. Available from: http://www.uzis.cz/en/fast-information/allergology-and-clinical-immunology-activity-branch-liberecky-region-2012 [Accessed 2015-06-16].

Kašková, R. (2013c): *Allergology and Clinical Immunology - Activity in Branch in the Pardubický Region in 2012.* [Online]. Available from: http://www.uzis.cz/e

n/fast-information/allergology-and-clinical-immunology-activity-branch-pardubicky-region-2012 [Accessed 2015-06-16].

Kašková, R. (2013d): *Allergology and Clinical Immunology - Activity in Branch in the Ústecký Region in 2012.* [Online]. Available from: http://www.uzis.cz/en/fast-information/allergology-and-clinical-immunology-activity-branch-ustecky-region-2012 [Accessed 2015-06-16].

Kozlíková, B. (2013a): *Activity in the branch allergology and clinical immunology in the Moravskoslezsky Region in 2012.* [Online]. Available from: http://www.uzis.cz/en/fast-information/activity-branch-allergology-and-clinical-immunology-moravskoslezsky-region-2012 [Accessed 2015-06-16].

Kozlíková, B. (2013b): *Activity in the branch allergology and clinical immunology in the Olomoucky Region in 2012.* Available from: http://www.uzis.cz/en/fast-information/activity-branch-allergology-and-clinical-immunology-olomoucky-region-2012 [Accessed 2015-06-16].

Lauridsen, J., Bech, M., López, F. & Sánchez, M. M. (2008): Geographic and Temporal Heterogeneity in Public Prescription Pharmaceutical Expenditures in Spain. *The Review of Regional Studies.* [Online] vol. 38, no. 1, p. 89-103. Available from: https://ideas.repec.org/a/rre/publsh/v38y2008i1p89-103.html [Accessed 2015-01-13]. ISSN 1553-0892.

LeSage, J. & Pace, R. K. (2009): *Introduction to Spatial Econometrics*. Statistics, Textbooks and monographs. Boca Raton: Chapman & Hall / CRC Press. ISBN 978-1-4200-6424-7.

Medical University of South Carolina (2015): *Allergy and Pesticides*. [Online]. Available from: http://academicdepartments.musc.edu/family_medicine/oem_agmed/Allergypesticides.htm [Accessed 2015-12-02].

Moran, P. A. P. (1948): The Interpretation of Statistical Maps. *Journal of the Royal Statistical Society.* [Online] Series B (Methodological), vol. 10, no. 2 (1948), pp. 243-251. Available from: http://www.jstor.org/stable/2983777 [Accessed: 2015-11-16].

Nagelkerke, N. J. D. (1991): A Note on a General Definition of the Coefficient of Determination. *Biometrika.* [Online] vol. 78, no. 3 (1991), pp. 691-692. Available from: http://www.jstor.org/stable/2337038 [Accessed 2015-11-25]. DOI 10.2307/2337038.

Nováková, Z. (2015): *FW: žádost o data pro diplomovou práci - 990.* [E-mail] Helena.Surova@uzis.cz. 1 June 2015.

OECD. (2014): *Geographic Variations in Health Care: What Do We Know and What Can Be Done to Improve Health System Performance?.* OECD Health Policy Studies. [Online] OECD Publishing. Available from: http://www.oecd-ilibrary.org/social-issues-migration-health/geographic-variations-in-health-care_9789264216594-en [Accessed 2015-01-13]. ISBN 978-92-64-21659-4.

Pazourková, J. (2013a): *Activity in Branch Allergology and Clinical Immunology in the Jihomoravský Region in 2012*. [Online]. Available from: http://www.uzis.cz/en/fast-information/activity-branch-allergology-and-clinical-immunology-jihomoravsky-region-2012 [Accessed 2015-06-16].

Pazourková, J. (2013b): *Activity in Branch Allergology and Clinical Immunology in the Vysočina Region in 2012.* [Online]. Available from: http://www.uzis.cz/en/fast-information/activity-branch-allergology-and-clinical-immunology-vysocina-region-2012 [Accessed 2015-06-16].

Pazourková, J. (2013c): *Activity in Branch Allergology and Clinical Immunology in the Zlinsky Region in 2012.* [Online]. Available from: http://www.uzis.cz/en/fast-information/activity-branch-allergology-and-clinical-immunology-zlinsky-region-2012 [Accessed 2015-06-16].

Rodriguez, V. & Dolejší, V. (2015): Rozdrobená medicína: Ministerstvo zjednoduší, jak se školí dětští lékaři nebo chirurgové. *Hospodářské noviny*. [Online] 3 November. Available from: http://archiv.ihned.cz/c1-64817070-rozdrobena-medicina-ministerstvo-zdravotnictvi-zjednodusi-zpusob-skoleni-na-specialisty [Accessed 2015-11-27].

Shaw, J. W., Horrace, W. C. & Vogel, R. J. (2005): The Determinants of Life Expectancy: An Analysis of the OECD Health Data. *Southern Economic Journal*. [Online], vol. 71, p. 768-783. Available from: http://faculty.maxwell.syr.edu/whorrace/publications/SEJ2005.pdf [Accessed 2015-03-09].

Šídlo, L. (2010): Regionální diferenciace věkové struktury praktických lékařů v České republice na počátku 21. století. *Praktický lékař*. [Online] vol. 90, no. 12. p 704-707. ISSN 0032-6739. Available from: http://www.prolekare.cz/praktick y-lekar-clanek/regionalni-diferenciace-vekove-struktury-praktickych-lekaru-v-ceske-republice-na-pocatku-21-stoleti-33549 [Accessed 2015-11-27].

Šídlo, L. & Tesárková, K. (2009): Současná regionální diferenciace lékařů primární péče v České republice z pohledu demografie. In *Sborník příspěvků z konference Reprodukce lidského kapitálu II – vzájemné vazby a souvislosti*. 15-16 December 2008, Prague: Department of Demography of Faculty of Informatics and Statistics of The University of Economics in Prague & Institute for Information on Education. [Online]. Available from: http://kdem.vse.cz/resources/relik09/Prispevky_PDF/Sidlo_Tesarkova.pdf [Accessed 2015-11-27].

Skinner, J. (2011): Causes and Consequences of Regional Variations in Health Care. *Handbook of Health Economics*. [Online] vol. 2, p. 45-93. Available from: https://www.dartmouth.edu/~jskinner/documents/Skinner_CausesandConsequences.pdf [Accessed 2015-12-04]. ISSN: 1574-0064.

Thode, H. C. (2002): *Testing for Normality*. STATISTICS: textbooks and monographs, vol. 164. New York: Marcel Dekker, Inc. ISBN 0-8247-9613-6.

Troch, T. (n.d.): *Datové rozhraní a metodika dat poskytovaných Ministerstvem zdravotnictví Univerzitě Karlově, Fakultě sociálních věd.* [Online]. Available from: http://ies.fsv.cuni.cz/cs/node/188 [Accessed 2015-06-16].

Tsai, C., Sullivan, A. F., Gordon, J. A., Kaushal, R., Magid, D. J., Blumenthal, D. & Camargo, C. A. (2009): Quality of care for acute asthma in 63 US emergency departments. *Journal of Allergy and Clinical Immunology*. [Online] vol. 123, issue 2, p. 354-361. Available from: http://www.sciencedirect.com/science/article/pii/S0091674908019301 [Accessed 2015-12-04]. DOI 10.1016/j.jaci.2008.10.051.

URRlab. (2013): *Historická data v GIS: Prostorová a statistická data ke stažení*. [Online]. Available from: http://www.historickygis.cz/cs/shp-vrstvy [Accessed 2015-11-03].

ÚZIS ČR. (2013): *Činnost zdravotnických zařízení ve vybraných oborech léčebně preventivní péče 2012*. [Online] Praha: Ústav zdravotnických informací a statistiky ČR. Available from: http://www.uzis.cz/katalog/zdravotnicka-statistika/cinnost-zdravotnickych-zarizeni-ve-vybranych-oborech [Accessed 2015-01-22]. ISBN 978-80-7472-063-5.

Všeobecná zdravotní pojišťovna České republiky (2010): *Smluvní odbornosti pracovišť.* [Online]. Available from: http://www.vzp.cz/poskytovatele/ciselniky/smluvni-odbornosti-pracovist [Accessed 2015-06-16].

Warner, J. O., Kaliner, M. A., Crisci, C. D., Del Giacco, S., Frew, A. J., Gh, L., Maspero, J., Moon, H., Nakagawa, T., Potter, P. C., Rosenwasser, L. J., Singh, A. B., Valovirta, E. & van Cauwenberge, P. (2006): Allergy Practice Worldwide: A Report by the World Allergy Organization Specialty and Training Council. *Allergy and Clinical Immunology International - World Allergy Organization Journal.* [Online], vol. 18, p. 4-10. Available from: http://www.researchgate.net/publication/7383553_Allergy_practice_worldwide_a_report_by_the_World_Allergy_Organization_Specialty_and_Training_Council [Accessed 2015-01-30]. DOI 10.1159/000090502.

Westert, G.P., Groenewegen, P.P., Boshuizen, H.C., Spreeuwenberg, P.M.M. & Steultjens, M.P.M. (2003): Medical practice variations in hospital care: time trends of a spatial phenomenon. Health & Place. [Online] vol. 10, issue 3, p. 215-220. Available from: http://www.sciencedirect.com/science/article/pii/S1353829203000674 [Accessed 2015-02-01]. DOI 0.1016/j.healthplace.2003.07.002.

Wooldridge, J. M. (2009): *Introductory Econometrics: A Modern Approach.* 4th ed. Mason: South-Western Cengage Learning, c2009. ISBN 978-0-324-66054-8.

Appendix A: Czech Districts

Table A.1: List of Czech districts and their LAU codes

LAU Code	District Name	LAU Code	District Name
CZ0100	Hl. m. Praha (Prague)	CZ0513	Liberec
CZ0201	Benešov	CZ0514	Semily
CZ0202	Beroun	CZ0521	Hradec Králové
CZ0203	Kladno	CZ0522	Jičín
CZ0204	Kolín	CZ0523	Náchod
CZ0205	Kutná Hora	CZ0524	Rychnov nad Kněžnou
CZ0206	Mělník	CZ0525	Trutnov
CZ0207	Mladá Boleslav	CZ0531	Chrudim
CZ0208	Nymburk	CZ0532	Pardubice
CZ0209	Praha-východ (Prague-east)	CZ0533	Svitavy
CZ020A	Praha-západ (Prague-west)	CZ0534	Ústí nad Orlicí
CZ020B	Příbram	CZ0631	Havlíčkův Brod
CZ020C	Rakovník	CZ0632	Jihlava
CZ0311	České Budějovice	CZ0633	Pelhřimov
CZ0312	Český Krumlov	CZ0634	Třebíč
CZ0313	Jindřichův Hradec	CZ0635	Žďár nad Sázavou
CZ0314	Písek	CZ0641	Blansko
CZ0315	Prachatice	CZ0642	Brno-město (Brno-town)
CZ0316	Strakonice	CZ0643	Brno-venkov (Brno-country)
CZ0317	Tábor	CZ0644	Břeclav
CZ0321	Domažlice	CZ0645	Hodonín
CZ0322	Klatovy	CZ0646	Vyškov
CZ0323	Plzeň-město (Plzeň-town)	CZ0647	Znojmo
CZ0324	Plzeň-jih (Plzeň-south)	CZ0711	Jeseník
CZ0325	Plzeň-sever (Plzeň-north)	CZ0712	Olomouc
CZ0326	Rokycany	CZ0713	Prostějov
CZ0327	Tachov	CZ0714	Přerov
CZ0411	Cheb	CZ0715	Šumperk
CZ0412	Karlovy Vary	CZ0721	Kroměříž
CZ0413	Sokolov	CZ0722	Uherské Hradiště
CZ0421	Děčín	CZ0723	Vsetín
CZ0422	Chomutov	CZ0724	Zlín
CZ0423	Litoměřice	CZ0801	Bruntál
CZ0424	Louny	CZ0802	Frýdek-Místek
CZ0425	Most	CZ0803	Karviná
CZ0426	Teplice	CZ0804	Nový Jičín
CZ0427	Ústí nad Labem	CZ0805	Opava
CZ0511	Česká Lípa	CZ0806	Ostrava-město (Ostrava-town)
CZ0512	Jablonec nad Nisou		

Source: Český statistický úřad (2014).

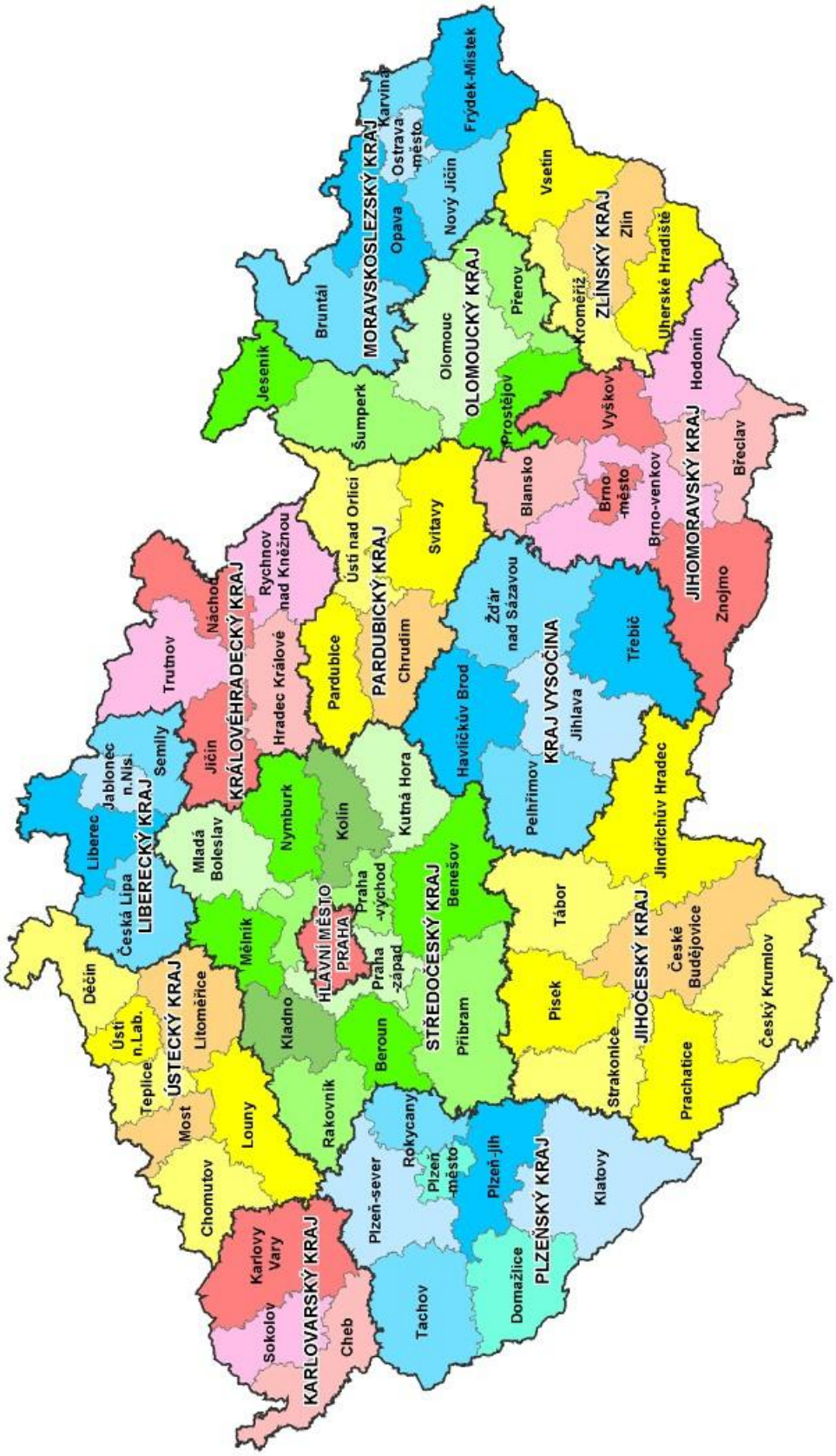

Figure A.1: Map of Czech regions and districts (2012)

Note: Regions are denoted by capitals. Prague (Hlavní město Praha) is both a district and a region.
Source: Czech Statistical Office (2012). Download file "Regions and districts in the CR - map" [JPG].

Appendix B: Alternative OLS and SAR Utilization Model Specifications

Table B.1: Care utilization - OLS with AvAge

Variable	Coefficient estimate	P-value for t statistic
$intercept$	6.990758	0.00000
PC_1	0.003682	0.76423
PC_2	- 0.041402	0.00360
PC_3	0.019521	0.40223
PC_4	0.051556	0.03138
$\log(PhysPC+1)$	0.317911	0.00002
$AvAge$	- 0.098294	0.00474
$\log(AvNeighPhysPC)$	0.210554	0.01150
F test of joint significance (p-value): $3.8 \cdot 10^{-8}$		
R^2: 0.4887, Adjusted R^2: 0.4368		

Source: Author's computations.

Table B.2: Care utilization - OLS with OAindex

Variable	Coefficient estimate	P-value for t statistic
$intercept$	3.620071	0.00000
PC_1	0.006051	0.62279
PC_2	- 0.043740	0.00231
PC_3	0.023823	0.31091
PC_4	0.052453	0.02871
$\log(PhysPC+1)$	0.320874	0.00002
$OAindex$	- 0.006153	0.00527
$\log(AvNeighPhysPC)$	0.214381	0.01023
F test of joint significance (p-value): $4.2 \cdot 10^{-8}$		
R^2: 0.4872, Adjusted R^2: 0.4352		

Source: Author's computations.

Table B.3: Care utilization - OLS with log of OAindex

Variable	Coefficient estimate	P-value for t statistic
$intercept$	5.932431	0.00000
PC_1	0.004916	0.68891
PC_2	- 0.042669	0.00282
PC_3	0.021186	0.36468
PC_4	0.050722	0.03442
$\log(PhysPC+1)$	0.323469	0.00002
$\log(OAindex)$	- 0.638511	0.00494
$\log(AvNeighPhysPC)$	0.217970	0.00903
F test of joint significance (p-value): $4.0 \cdot 10^{-8}$		
R^2: 0.4881, Adjusted R^2: 0.4362		

Source: Author's computations.

Table B.4: Care utilization - OLS with AvNeighPhysPC

Variable	Coefficient estimate	P-value for t statistic
$intercept$	18.060919	0.00074
PC_1	0.004043	0.74387
PC_2	- 0.041673	0.00363
PC_3	0.015997	0.49326
PC_4	0.052554	0.03023
$\log(PhysPC+1)$	0.318361	0.00002
$\log(AvAge)$	- 4.062839	0.00442
$AvNeighPhysPC$	0.068902	0.02108
F test of joint significance (p-value): $6.4 \cdot 10^{-8}$		
R^2: 0.4804, Adjusted R^2: 0.4277		

Source: Author's computations.

Table B.5: Care utilization - OLS with log of NeighPhysPC

Variable	Coefficient estimate	P-value for t statistic
$intercept$	15.774607	0.00328
PC_1	0.003288	0.78949
PC_2	- 0.040048	0.00497
PC_3	0.015317	0.50905
PC_4	0.051505	0.03250
$\log(PhysPC+1)$	0.336828	0.00001
$\log(AvAge)$	- 3.451987	0.01574
$\log(NeighPhysPC)$	0.178992	0.01414
F test of joint significance (p-value): $4.6 \cdot 10^{-8}$		
R^2: 0.4857, Adjusted R^2: 0.4336		

Source: Author's computations.

Table B.6: Care utilization - OLS with NeighPhysPC

Variable	Coefficient estimate	P-value for t statistic
intercept	15.251123	0.00525
PC_1	0.002731	0.82599
PC_2	- 0.038810	0.00704
PC_3	0.010190	0.66188
PC_4	0.052587	0.03123
log(PhysPC+1)	0.338546	0.00001
log(AvAge)	- 3.303703	0.02320
NeighPhysPC	0.051593	0.02581
F test of joint significance (p-value): $7.5 \cdot 10^{-8}$		
R^2: 0.4778, Adjusted R^2: 0.4248		

Source: Author's computations.

Table B.7: Care utilization - OLS with Pdens

Variable	Coefficient estimate	P-value for t statistic
$intercept$	18.58748	0.00055
PC_1	-0.02195	0.43129
PC_2	-0.04401	0.00246
PC_3	-0.03124	0.56758
PC_4	0.05008	0.03665
log($PhysPC$+1)	0.31676	0.00002
log($AvAge$)	-4.22841	0.00328
log($AvNeighPhysPC$)	0.23773	0.00678
$Pdens$	0.02505	0.30934
F test of joint significance (p-value): $7.7 \cdot 10^{-8}$		
R^2: 0.4958, Adjusted R^2: 0.4362		

Source: Author's computations.

Table B.8: Care utilization - OLS with log of Pdens

Variable	Coefficient estimate	P-value for t statistic
$intercept$	19.00753	0.00040
PC_1	0.03981	0.16306
PC_2	- 0.03735	0.00906
PC_3	0.06341	0.10557
PC_4	0.05747	0.01801
$\log(PhysPC+1)$	0.36374	0.00001
$\log(AvAge)$	- 4.34256	0.00246
$\log(AvNeighPhysPC)$	0.25593	0.00422
$\log(Pdens)$	- 0.14369	0.15927
F test of joint significance (p-value): $5.0 \cdot 10^{-8}$		
R^2: 0.4964, Adjusted R^2: 0.4366		

Source: Author's computations.

Table B.9: Care utilization - OLS with Unemp

Variable	Coefficient estimate	P-value for t statistic
$intercept$	17.548442	0.00096
PC_1	0.003367	0.78409
PC_2	- 0.047015	0.00256
PC_3	0.024864	0.30229
PC_4	0.060564	0.02003
$\log(PhysPC+1)$	0.299933	0.00010
$\log(AvAge)$	- 3.944854	0.00536
$\log(AvNeighPhysPC)$	0.192988	0.02357
$Unemp$	0.014724	0.34269
F test of joint significance (p-value): $8.2 \cdot 10^{-8}$		
R^2: 0.4955, Adjusted R^2: 0.4361		

Source: Author's computations.

Table B.10: Care utilization - OLS with log of Unemp

Variable	Coefficient estimate	P-value for t statistic
$intercept$	17.930944	0.00075
PC_1	0.004104	0.73883
PC_2	- 0.046325	0.00261
PC_3	0.025155	0.29981
PC_4	0.059694	0.02083
$\log(PhysPC+1)$	0.299786	0.00010
$\log(AvAge)$	- 4.068978	0.00419
$\log(AvNeighPhysPC)$	0.195371	0.02142
$\log(Unemp)$	0.094312	0.36139
F test of joint significance (p-value): $8.4 \cdot 10^{-8}$		
R^2: 0.4950, Adjusted R^2: 0.4356		

Source: Author's computations.

Table B.11: Care utilization - SAR with AvAge

Variable	Coefficient estimate	P-value of the Wald test
$intercept$	4.98515	0.000168
PC_1	- 0.00272	0.804343
PC_2	- 0.02661	0.032109
PC_3	0.00127	0.950861
PC_4	0.05380	0.008638
$\log(PhysPC+1)$	0.26797	0.000019
$AvAge$	- 0.08109	0.007272
$W\log(UtilPC)$	0.44097	0.000046
Log-likelihood: 15.48423		
AIC: -12.968		

Source: Author's computations.

Table B.12: Care utilization - SAR with log of OAindex

Variable	Coefficient estimate	P-value of the Wald test
intercept	4.10168	0.000058
PC_1	- 0.00184	0.867044
PC_2	- 0.02738	0.027238
PC_3	0.00224	0.913220
PC_4	0.05327	0.009255
log(PhysPC+1)	0.27192	0.000016
log(OAindex)	- 0.53007	0.007060
W log(UtilPC)	0.45027	0.000026
Log-likelihood: 15.48958		
AIC: -12.979		

Source: Author's computations.

Table B.13: Care utilization - SAR with OAindex

Variable	Coefficient estimate	P-value of the Wald test
$intercept$	2.19065	0.000001
PC_1	- 0.00084	0.938783
PC_2	- 0.02837	0.022698
PC_3	0.00460	0.823693
PC_4	0.05459	0.007694
$\log(PhysPC+1)$	0.27004	0.000018
$OAindex$	- 0.00510	0.007822
$W\log(UtilPC)$	0.44650	0.000035
Log-likelihood: 15.44687		
AIC: -12.894		

Source: Author's computations.

Table B.14: Care utilization - SAR with Pdens

Variable	Coefficient estimate	P-value of the Wald test
intercept	14.09852	0.002460
PC_1	- 0.00885	0.709227
PC_2	- 0.02712	0.032069
PC_3	- 0.01119	0.812959
PC_4	0.05390	0.008647
$\log(PhysPC+1)$	0.26778	0.000019
$\log(AvAge)$	- 3.35381	0.006991
Pdens	0.00590	0.776139
$W\log(UtilPC)$	0.44235	0.000046
Log-likelihood: 15.52996		
AIC: -11.060		

Source: Author's computations.

Table B.15: Care utilization - SAR with log of Pdens

Variable	Coefficient estimate	P-value of the Wald test
intercept	14.61383	0.001337
PC_1	0.02952	0.209778
PC_2	- 0.02143	0.083533
PC_3	0.03855	0.223349
PC_4	0.06129	0.003489
log(PhysPC+1)	0.30383	0.000004
log(AvAge)	- 3.54116	0.003650
log(Pdens)	- 0.13078	0.115114
W log(UtilPC)	0.48985	0.000003
Log-likelihood: 15.52893		
AIC: -13.058		

Source: Author's computations.

Table B.16: Care utilization - SAR with Unemp

Variable	Coefficient estimate	P-value of the Wald test
intercept	13.70714	0.002743
PC_1	- 0.00289	0.790765
PC_2	- 0.03349	0.012303
PC_3	0.00863	0.683698
PC_4	0.06277	0.003572
log(PhysPC+1)	0.24887	0.000105
log(AvAge)	- 3.25609	0.007591
Unemp	0.01698	0.204485
W log(UtilPC)	0.42732	0.000067
Log-likelihood: 16.29354		
AIC: -12.587		

Source: Author's computations.

Table B.17: Care utilization - SAR with log of Unemp

Variable	Coefficient estimate	P-value of the Wald test
intercept	14.13357	0.002035
PC_1	- 0.00206	0.850064
PC_2	- 0.03263	0.013579
PC_3	0.00880	0.679883
PC_4	0.06195	0.003874
log(PhysPC+1)	0.24848	0.000116
log(AvAge)	- 3.39630	0.005472
log(Unemp)	0.10852	0.225049
W log(UtilPC)	0.42938	0.000062
Log-likelihood: 16.22458		
AIC: -12.449		

Source: Author's computations.