- 733 Removals + 810 Additions

Massaroetal.
RESEARCH
Leveraging insurance customer data to
characterize socioeconomic indicators of Swiss
municipalities
Emanuele Massaro
1*
, Lorenzo Donadio
1
,ClaudiaR.Binder
1
and Rossano Schifanella
2,3
*
Correspondence:
emanuele.massaro@epfl.ch
1
HERUS Lab,
´
Ecole polytechnique
f´ed´erale de Lausanne, ENAC, IIE
GR C1 455 (Bˆatiment GR) –
Station 2, 1015 – Lausanne, CH
Full list of author information is
available at the end of the article
Abstract
In this paper, we explore the potential for insurance customers data to predict
socio-economic indicators of Swiss municipalities. First, we create a feature space
by defining various meaningful characteristics of a particular city through the
aggregation of individual data. We then evaluate those quantities by considering
available ocial statistics for Swiss municipalities for di↵erent categories, i.e.,
Population, Transportation, Work, Space and Territory, Housing, and Economy,
and we investigate whether they can be predicted based on our feature space. For
the purpose of prediction, our study adopts two geographical regression models
to explore the impact of spatial dependency e↵ects. Results show consistently a
correlation between customers' individual characteristics and ocial
socio-economic indexes. Performance fluctuates depending on the category,
reaching an adjustedR
2
=0.6in the case of per inhabitant housing area using a
5-fold cross validation setting. As a case study, we discuss in detail the
predictability of the percentage of the population using public transportation and

Donadioetal.
RESEARCH
Leveraging insurance customer data to
characterize socioeconomic indicators of Swiss
municipalities
Lorenzo Donadio
1
, Rossano Schifanella
2,3
,ClaudiaR.Binder
1
and Emanuele Massaro
1*
*
Correspondence:
ema.massaro@gmail.com
1
HERUS Lab,
´
Ecole polytechnique
f´ed´erale de Lausanne, ENAC, IIE
GR C1 455 (Bˆatiment GR) –
Station 2, 1015 – Lausanne, CH
Full list of author information is
available at the end of the article
Abstract
In this paper, we explore the potential for insurance customers data to predict
socio-economic indicators of Swiss municipalities. First, we create a feature space
by defining various meaningful characteristics of a particular city through the
aggregation of individual data. We then evaluate those quantities by considering
available ocial statistics for Swiss municipalities for di↵erent categories, i.e.,
Population, Transportation, Work, Space and Territory, Housing, and Economy,
and we investigate whether they can be predicted based on our feature space. For
the purpose of prediction, our study adopts two geographical regression models
to explore the impact of spatial dependency e↵ects. Results show consistently a
correlation between customers' individual characteristics and ocial
socio-economic indexes. Performance fluctuates depending on the category,
reaching an adjustedR
2
=0.6in the case of per inhabitant housing area using a
5-fold cross validation setting. As a case study, we discuss in detail the
predictability of the percentage of the population using public transportation and

- 733 Removals + 810 Additions

for analyzing and predicting the impact of human activity on
urban well being on spatial and temporal scales that are no consistent with the
ocial statistics available (e.g., small communes, districts or neighborhoods).
Keywords:insurance data; geographical regression; socioeconomic indicators;
Swiss municipalities
1 Introduction
National Statistical Institutes (NSIs) play an important role in modern societies
to release precise information on social, environmental or economical activities [1]
in the form of a census. For example, the census records key aspects such as how
many people live in an area, their ages and their income per capita, and it en-
ables the prediction of future population to inform the need for schools, homes or
public services. Censuses are essential for many of the indicators that enable us to
measure progress towards the Sustainable Development Goals [2]. Ocial statistics
on socioeconomic status are increasingly addressing a significant modernization of
their production processes, nationally and internationally [3]. This is also due to the
opportunities o↵ered by the use of new data sources, such as mobile phone data [4],
social media [5], satellite images [6], credit card transactions [7] and others [8,9,10].
The goal of NSIs is to integrate and combine this new information with the tra-
ditional sources of investigations and administrative archives, and the increasingly

Massaroetal.Page 2 of29
widespread orientation towards the construction of registers of elementary data inte-
grated. Three important challenges rise: i) the quality and methodology of the data
collection, ii) privacy and legal issues and iii) the processing, storage and transfer
of large data sets. Data sources such as social media, and mobile phone records,
do not have a well-defined target population, structure and quality (see Section2
for a literature review) that make dicult to apply traditional statistical methods
based on sampling theory. Privacy and legal aspects pose another challenge: the
prevention of the disclosure of the identity of individuals is regulated and enforced
by international laws, and ensuring an appropriate level of privacy is challenging
in case of heterogeneous, and multi-source large scale data streams [11]. Copyright

open new directions
for analyzing and predicting the impact of human activity on
urban well being on spatial and temporal scales that are no consistent with the
ocial statistics available (e.g., small communes).
Keywords:insurance data; geographical regression; socioeconomic indicators;
Swiss municipalities
1 Introduction
National Statistical Institutes (NSIs) play an important role in modern societies
to release precise information on social, environmental or economical activities [1]
in the form of a census. For example, the census records key aspects such as how
many people live in an area, their ages and their income per capita, and it en-
ables the prediction of future population to inform the need for schools, homes or
public services. Censuses are essential for many of the indicators that enable us to
measure progress towards the Sustainable Development Goals [2]. Ocial statistics
on socioeconomic status are increasingly addressing a significant modernization of
their production processes, nationally and internationally [3]. This is also due to the
opportunities o↵ered by the use of new data sources, such as mobile phone data [4],
social media [5], satellite images [6], credit card transactions [7] and others [8,9,10].
The goal of NSIs is to integrate and combine this new information with the tra-
ditional sources of investigations and administrative archives, and the increasingly

Donadioetal.Page 2 of30
widespread orientation towards the construction of registers of elementary data inte-
grated. Three important challenges rise: i) the quality and methodology of the data
collection, ii) privacy and legal issues and iii) the processing, storage and transfer
of large data sets. Data sources such as social media, and mobile phone records,
do not have a well-defined target population, structure and quality (see Section2
for a literature review) that make dicult to apply traditional statistical methods
based on sampling theory. Privacy and legal aspects pose another challenge: the
prevention of the disclosure of the identity of individuals is regulated and enforced
by international laws, and ensuring an appropriate level of privacy is challenging
in case of heterogeneous, and multi-source large scale data streams [11]. Copyright

the storage and the transfer of large amount of heterogeneous information ensur- ing security [13]. In this context, countries are increasingly favouring alternative means of gathering information, instead of thetraditionaltechniques of sending out printed forms, interviewing people in person, or via the use of online questionnaires. Alternatively, they are looking to indirect means of collecting data, taking advan- tage of a wide spectrum of administrative data streams that act as a proxy for the variables of interest. In this direction, customers insurance records represent a valuable input to model the socioeconomic substrate of cities, and an opportunity for policy makers and researchers to broaden the scope of their studies. Social scientists raised the issue of representativeness and sampling bias of large scale digital data. For example, in [14] the authors show how age, gender, ethnicity, socioeconomic status, online ex- periences, and Internet skills, influence the social network sites that users generally adopt. This has implications for the extent of the conclusions that a study could claim given a particular audience. Like census data, insurance customers records share a similar size, reliability, and structural complexity [15]. However, they di↵er in their spatio-temporal granularity and collection costs. In fact, the information of insurance customers is collected constantly by the provider while the census runs generally with a multi-yearly frequency due to its organizational costs. A downside is the proprietary nature of customers records that could invalidate the possible benefits for a broader community. However, we embrace the vision of initiatives likeData Collaboratives
[1]
that propose a new form of collaboration, beyond the public-private partnership model, in which participants from di↵erent sectors, in particular companies, exchange their data to create public value. In this research, we develop a methodology to predict socioeconomic indicators at a city level using individual customers data from an insurance company in Switzerland. Swiss mu- nicipalities, sometimes also calledcommunities, are the lowest administrative level in the country. The responsibilities of the 2.212 (as of 1 January 2019) Swiss mu- nicipalities is decided by each Cantons. These may include the provision of local public services such as education, medical and social assistance, public transport

the storage and the transfer of large amount of heterogeneous information ensur- ing security [13]. In this context, countries are increasingly favouring alternative means of gathering information, instead of thetraditionaltechniques of sending out printed forms, interviewing people in person, or via the use of online questionnaires. Alternatively, they are looking to indirect means of collecting data, taking advan- tage of a wide spectrum of administrative data streams that act as a proxy for the variables of interest. In this direction, customers insurance records represent a valuable input to model the socioeconomic substrate of cities, and an opportunity for policy makers and researchers to broaden the scope of their studies. Social scientists raised the issue of representativeness and sampling bias of large scale digital data. For example, in [14] the authors show how age, gender, ethnicity, socioeconomic status, online ex- periences, and Internet skills, influence the social network sites that users generally adopt. This has implications for the extent of the conclusions that a study could claim given a particular audience. Like census data, insurance customers records share a similar size, reliability, and structural complexity [15]. However, they di↵er in their spatio-temporal granularity and collection costs. In fact, the information of insurance customers is collected constantly by the provider while the census runs generally with a multi-yearly frequency due to its organizational costs. A downside is the proprietary nature of customers records that could invalidate the possible benefits for a broader community. However, we embrace the vision of initiatives likeData Collaboratives
[1]
that propose a new form of collaboration, beyond the public-private partnership model, in which participants from di↵erent sectors, in particular companies, exchange their data to create public value. In this research, we develop a methodology to predict socioeconomic indicators at a city level using individual customers data from an insurance company in Switzerland. Swiss mu- nicipalities, sometimes also calledcommunities, are the lowest administrative level in the country. The responsibilities of the 2.212 (as of 1 January 2019) Swiss mu- nicipalities is decided by each Cantons. These may include the provision of local public services such as education, medical and social assistance, public transport

- 733 Removals + 810 Additions

ecutive power) and by the general assembly of all adult Swiss residents (legislative
[1]
↵

Massaroetal.Page 3 of29
power). Many cantons leave the larger municipalities the option of opting for a city
parliament. Swiss citizenship is based on the citizenship of a municipality. Every
Swiss citizen is, first, a citizen of a municipality (right of citizenship of the city or
of origin) and, then, of a canton (right of cantonal or indigenous citizenship). For
all these reasons, our analysis adopts the municipality as a spatial unit of reference.
We propose a two-steps process to predict a wide range of socioeconomic indi-
cators. First, we compute a set of behavioral metrics using customers activity logs
concerning housing properties and vehicles insured by “la Mobili`ere” in 2017. Sec-
ond, we use those microeconomics indicators as explanatory variables on two spatial
regression models to predict 12 socioeconomic indices of 170 Swiss municipalities
for which we have reliable ocial statistical data as ground truth. In this work, we
focus on indices related to six di↵erent categories, i.e., Population, Transportation,
Work, Space and Territory, Housing, and Economy. We show that insurance data
customers can represent a valid resource to model socioeconomic indicators at scale.
The rest of the paper is organized as follows. Section2describes how insurance
data can benefit the urban data science research and it provides an overview of
the previous work in this area. Section3provides insights into the two datasets
that we use in this paper: insurance and census data. In Section4we describe the
methodology and the modeling framework. Section6discusses the results presented
in Section5and it provides a critical view on the limitations of the implemented
approach. Finally, in Section7we summarize the importance and the findings of
this research and we provide insights for future directions.
2 Related Work
Researchers across various disciplines including sociology, demography and public
health have been keen on examining how society functions observing
populations at
scale. Socio-economic indicators of cities, which were investigated before the
digital

ecutive power) and by the general assembly of all adult Swiss residents (legislative
[1]
https://datacollaboratives.org

Donadioetal.Page 3 of30
power). Many cantons leave the larger municipalities the option of opting for a city
parliament. Swiss citizenship is based on the citizenship of a municipality. Every
Swiss citizen is, first, a citizen of a municipality (right of citizenship of the city or
of origin) and, then, of a canton (right of cantonal or indigenous citizenship). For
all these reasons, our analysis adopts the municipality as a spatial unit of reference.
We propose a two-steps process to predict a wide range of socioeconomic indi-
cators. First, we compute a set of behavioral metrics using customers activity logs
concerning housing properties and vehicles insured by “la Mobili`ere” in 2017. Sec-
ond, we use those microeconomics indicators as explanatory variables on two spatial
regression models to predict 12 socioeconomic indices of 170 Swiss municipalities
for which we have reliable ocial statistical data as ground truth. In this work, we
focus on indices related to six di↵erent categories, i.e., Population, Transportation,
Work, Space and Territory, Housing, and Economy. We show that insurance data
customers can represent a valid resource to model socioeconomic indicators at scale.
The rest of the paper is organized as follows. Section2describes how insurance
data can benefit the urban data science research and it provides an overview of
the previous work in this area. Section3provides insights into the two datasets
that we use in this paper: insurance and census data. In Section4we describe the
methodology and the modeling framework. Section6discusses the results presented
in Section5and it provides a critical view on the limitations of the implemented
approach. Finally, in Section7we summarize the importance and the findings of
this research and we provide insights for future directions.
2 Related Work
Researchers across various disciplines including sociology, demography and public
health have been keen on examining how society functions observing
populations
at scale. Socioeconomic indicators of cities, which were investigated before the
dig-
ital

733 Removals 810 Additions

finding correlations between demographic factors
di↵erences
145 between
urban and suburban areas [16], crime rate [18], p
opulation health [19],
res-
146 idential
segregation [20] or waste production [17].
Since the aforementioned findings
147 were mostly based on survey results, they
may have been a↵ected by the fact that
148 people could have altered their answers
knowing that they weremonitored.Today
in the digital era in some cases information abou
149 t people behavior is collected even
without them being aware of that, let alone with
150 their informed consent.
In the era where the usage of digital technologie
151 s is so omnipresent, people every
152 day leave more digital trails than
we are currently able to process. Indeed, most of
153 the recent studies focused on the use
of digital andbig datato predict and study
154 socio-economic
indicators of cities and countries. In
extensive amount of related
155 work
scholars used di↵erent digital datasets for di↵er
ent research purposes such as
156 for studying human dynamics through cell phone
data [4,21], social media posts [22],
157 vehicle GPS traces [23] or credit card
data [7,24].
158 It has been shown that social media data
can predict the interplay between de-
159 mographic attributes and gender gap [25], the
monitoring [26] and assimilation of
160 migrants [27]), unemployment [28], and health
outcomes [9]. Social media are in-
161 creasingly used for demographic attribute
prediction [29]. Using web search engine
162
163 Massaroetal.Page 4 of29
164 Figure 1: Comparison between the spatial coverage of the insurance and the Union
165 of Swiss Cities datasets. (A) Total number of ins urance customers at zip code level.
166 The total number of customers is 1,341,328 which represents nearly 15% of the
167 Swiss population. (B) Number of inhabitants aggre gated at municipality level. The
168 total surface covered by those municipalities is almost 3890km
169 2
170 which represents
171 9.35% of the country’s area.

finding correlations between demographic factors
di↵erences in
urban and suburban areas [16], crime rate [18], p
146 opulation health [19],
residential
147 segregation [20] or waste production [17].
In all these cases the results were based
on the active participation of individuals to sur
148 veys, that
may have been a↵ected
149 by the tendency of respondents to alter their
behavior knowing that they aremon-
itored. On the contrary, the digitalization of th
150 e modern society allows to model
human behavior by means of indirect and less intr
151 usive data collection methodolo-
gies, that are often a byproduct of services desi
152 gned for di↵erent purposes. In a
scenario where the collectivity produces every da
153 y
more digital footprints than we
154 are able to process, the majority of
the recent studies focused on the use ofbig data
155 streams to predict and study socioeconomic
indicators of cities and countries. An
156 extensive body of work
adopt digital traces such as cell phone records
[4,21], social
157 media posts [22],
vehicle GPS traces [23] or credit card
transactions [7,24]tomodel
158 human dynamics at scale. For example, it
has been shown that social media data
159 can predict the interplay between demographic
attributes and gender gap [25], the
160 monitoring [26] and assimilation of
migrants [27]), unemployment [28], and health
161 outcomes [9]. Social media are increasingly
used for demographic attribute predic-
162 tion [29]. Using web search engine
datasets, Weber and Castillo [30] inferred gende
r,
163
164 Donadioetal.Page 4 of30
165 Figure 1: Comparison between the spatial coverage of the insurance and the Union
166 of Swiss Cities datasets. (A) Total number of ins urance customers at zip code level.
167 The total number of customers is 1,341,328 which represents nearly 15% of the
168 Swiss population. (B) Number of inhabitants aggre gated at municipality level. The
169 total surface covered by those municipalities is almost 3890km
170 2
171 which represents
172 9.35% of the country’s area.

733 Removals 810 Additions

173 ber and Jaimes [31]
analyzed data rom di↵erent ZIP codes enriched by
US census
174 data and exploited them to
highlight di↵erences in user behaviour
and search pat-
175 terns among several demographic
groups. Gender and age can also be inferred using
176 call detail records from smartphone
devices over large populations.
177 Our work belongs to this line of work, however, i
t explores for the first time,
178 to the best of our knowledge, the use of insuranc
e customer records to predict
179 census variables. Insurance data have been mostly
used to study the impact of
180 specific diseases [32,33], to propose models of
customers' fraud detection [34,35],
181 to understand
the correlation between census-based socioeconom
ic indicators
for
182 and injury
causes [36] or to evaluate disparities within hea
lth care systems [37].
183 3 Data
184 In this work, we make use of two main
datasets: 1) the customer activity logs of a
185 Swiss
insurance company and 2) the ocial socioeconomic
indicators of all the Swiss
186 municipalities with more than 10,000 inhabitants,
collected by the initiative called
187 Union des villes Suisse
188 [2]
, i.e. Union of Swiss Cities that has published s
189 tatistics
190 on 173 Swiss municipalities every year since
2006: in this research we use data
191 related to the year 2017 [38].
The two datasets have di↵erent spatial
aggregations:
192 while the information on the
insurance customers is at the
zip code level; the socio-
193 economic indicators have been collected at the
municipality level. In aggregating
the insurance data to match the spatial granulari
194 ty of the municipality data, we
195 restrict our
analysis to the 170 municipalities that are prese
nt in both
datasets.
196 Figure1compares

174 highlight di↵erences in navigational and search
patterns among several demographic
175 groups. Gender and age are also inferred using
call detail records from smartphone
176 devices over large populations.
177 Our work belongs to this line of work, however, i
t explores for the first time,
178 to the best of our knowledge, the use of insuranc
e customer records to predict
179 census variables. Insurance data have been mostly
used to study the impact of
180 specific diseases [32,33], to propose models of
customers fraud detection [34,35], to
181 explore
the correlation between census-based socioeconom
ic indicators
and injury
causes [36] or to evaluate disparities within hea
182 lth care systems [37].
183 3 Data
184 In this work, we make use of two
datasets: 1) the customer activity logs of
a Swiss
insurance company and 2) the ocial socioeconomic
185 indicators of all the Swiss
186 municipalities with more than 10,000 inhabitants,
collected by the initiative called
187 Union des villes Suisse
188 [2]
189 that publishes statistics
on 173 Swiss municipalities every
190 year since 2006. In this research, we
focus on a cross-sectional snapshot for 2017
[38].
191 The two datasets have di↵erent spatial
aggregations: while the information on the
192 insurance customers is at zip code level; the
socioeconomic indicators have been
193 collected at municipality level.
To match the spatial granularity, we restrict our
analysis to the 170 municipalities that are prese
194 nt in both
datasets. Figure1com-
195 pares
the spatial coverage in the two cases, while
Table1highlights di↵erences and
196

The dataset contains the housing and vehicles insurance policies of 1,341,328
anonymized customers of La Mobili`ere who were active
in 2017. La Mobili`ere is
[2]
https://uniondesvilles.ch/

Massaroetal.Page 5 of29
Table 1: Comparison between the two datasets.
Insurance DataUnion of Swiss Cities
FrequencyEvery yearEvery year
Spatial aggregationZip codeMunicipality
Data points3,185173
CostNot expensiveExpensive
DesignFor insurance marketingFor statistical analysis
AvailabilityPrivatePublic
a Swiss insurance group (brands: Die Mobiliar, La Mobili`ere, La Mobiliare) that is
organized as a holding company headed by a cooperative. The company was founded
in 1826, making it the oldest private insurance company in Switzerland. Mobiliar
is an all–insurance company operating exclusively in Switzerland and Liechtenstein.
With a market share of over 29%, it is the leader in the personal property insurance
market. Customers' details are aggregated at the level of the 170 municipalities for
which we have ocial census data using the postal code of provenance; this step
leaves us with 568,426 customers matching the geographical boundaries. For each
user, we have three types of information: i)demographic, e.g., age, gender, zip code
of the residential area, employment status, civil status; ii)car insurance, e.g., how
many cars are insured, the type, brand and price of the vehicles, as well as the
record of claims in terms of frequency and cost; iii)housing insurance, e.g., the
number of private buildings or houses insured with the company, the price of the
building and the logs of the claims. Table2summarizes the information available.
From this complete set of variables, we perform a feature engineering step in which
we select the variables of interest with the aid of a domain expert. As a result of
this process, we end up with 34 features as summarized in Table3.Werefertothis
feature space in the modeling section of this work.
3.2 Swiss Census Data
The ocial statistics for the Swiss municipalities

The dataset contains the housing and vehicles insurance policies of 1,341,328
anonymized customers of La Mobili`ere who were active
during 2017. La Mobili`ere
is a Swiss insurance group (brands: Die Mobiliar, La Mobili`ere, La Mobiliare) that
is organized as a holding company headed by a cooperative. The company was
[2]
https://uniondesvilles.ch/

Donadioetal.Page 5 of30
Table 1: Comparison between the two datasets.
Insurance DataUnion of Swiss Cities
FrequencyEvery yearEvery year
Spatial aggregationZip codeMunicipality
Data points3,185173
CostNot expensiveExpensive
DesignFor insurance marketingFor statistical analysis
AvailabilityPrivatePublic

- 733 Removals + 810 Additions

[3]
.
The report is published in the first quarter of the year and it presents varied facets
of the urban life; we focus on six domains:population,transportation,employment,
space and territory,housingandeconomy. We collect a total of 86 features for each
municipality: 11 indicators for transportation (t), 29 for population (p), 11 for em-
ployment (w), 8 for space and territory (s), 18 for housing (h) and 9 for economy
(e). From the original dataset we focus on the key target variables that are not
redundant and can be a proxy for quality of life in cities, such as the unemploy-
ment rate [24], use of public transportation [39] or investment in culture [40]. The
complete list of selected target variables is listed below. As a result, we select two
characteristics for each domain: fraction of foreigners and rate of beneficiaries of
social assistance (p), number of private cars per person and the fraction of com-
muters using public transportation (t), unemployment rate and the unemployment
rate among women (w), percentage of areas covered by buildings or green areas (s),
[3]
https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/publications/
ouvrages-synthese/statistiques-villes-suisses.html

Massaroetal.Page 6 of29
Table 2: Information for each customers in the insurance dataset
CatergoryVariable NameDescriptionVariable Type
DemographicNmbrAnonymous IDAlphanumeric
JobStateEmployement statusString
CivilCivil StatusString
GenderGenderString
YearOfBirthYear of birthInteger
Own/RentIf own or rent an houseBoolean (Yes/No)
LangSpeaking languageString (French, German,
Italian)
NationNation of originString
ZIPZip code of residence5-digit code
Children0-26How many childrenInteger
CarsCar1Canton*Canton where the car is registered String
Car1Brand*Brand of the carString
Car1Price*Price of the car (CHF)Integer
Car1ccm*Cylinder capacityInteger
Car1ClaimsCt5Y*Number of claims over the last 5
years
Integer
Car1ClaimsSum5Y*Sum of the money from the claims
over the last 5 years (CHF)
Float

Table 2: Information for each customers in the insurance dataset
CatergoryVariable NameDescriptionVariable Type
DemographicNmbrAnonymous IDAlphanumeric
JobStateEmployement statusString
CivilCivil StatusString
GenderGenderString
YearOfBirthYear of birthInteger
Own/RentIf own or rent an houseBoolean (Yes/No)
LangSpeaking languageString (French, German,
Italian)
NationNation of originString
ZIPZip code of residence5-digit code
Children0-26How many childrenInteger
CarsCar1Canton*Canton where the car is registered String
Car1Brand*Brand of the carString
Car1Price*Price of the car (CHF)Integer
Car1ccm*Cylinder capacityInteger
Car1ClaimsCt5Y*Number of claims over the last 5
years
Integer
Car1ClaimsSum5Y*Sum of the money from the claims
over the last 5 years (CHF)
Float

- 733 Removals + 810 Additions

Integer
StandoffurnStandard of furnitureInteger (1–2–3–4–5)
RoomsNumber of roomsInteger
BuildZipZip Code of the insured building5-Digit code
BuildInsSumTotal sum of the insured values of
the building (CHF)
Integer
YearofcontrsYear of constructions of the building 4-Digit Integer
TypeType of buildingString
HHaBClaimsCt5YNumber of claims over the last 5
years
Integer
HHaBClaimsSum5YSum of the money from the claims
over the last 5 years (CHF)
Integer
HHandBldPremPremium class of the buildingString
vacancy rate and average area per inhabitant (h), and municipal debt and fraction
of investment in culture (e). The complete list of the selected features is reported

in Table4.
Validation.As a validation step, we test the representativeness of the insurance
data along four dimensions: (a) total population, (b) percentage of foreigners, (c)
percentage of population aged 20-40, and (d) percentage of population aged 0-19.

Figure2presents the Pearson's correlation between these variables extracted from

the ocial census data and the La Mobili`ere customers base respectively. A high

degree of correlation (⇢=0.91) can be observed for the total population, meaning
that the insurance dataset mimics quite well the population distribution at the
municipality level. Focusing on the age facet, we observe a strong relation for
the
case of
customers in the age range of 20-40 (⇢=0.8) while the correlation disappears
(⇢=0.05) for customers aged 0-19. This behavior is expected since children and
teenagers are not usually the owners of insurance policies on vehicles or houses.
Last, we observe a solid relation with the
percentage of foreigners (⇢=0.6).
Spatial dependency.An aspect worth mentioning is that the target variables we
are interested in are embedded in space. In this direction, we compute the Moran's
I coecient to test spatial dependencies and, in p

Integer
StandoffurnStandard of furnitureInteger (1–2–3–4–5)
RoomsNumber of roomsInteger
BuildZipZip Code of the insured building5-Digit code
BuildInsSumTotal sum of the insured values of
the building (CHF)
Integer
YearofcontrsYear of constructions of the building 4-Digit Integer
TypeType of buildingString
HHaBClaimsCt5YNumber of claims over the last 5
years
Integer
HHaBClaimsSum5YSum of the money from the claims
over the last 5 years (CHF)
Integer
HHandBldPremPremium class of the buildingString
founded in 1826, making it the oldest private insurance company in Switzerland.
With a market share of over 29%, it is the leader in the personal property insurance
market. For each user, we have three classes of information: i)demographic, e.g.,
age, gender, zip code of the residential area, employment and civil status; ii)cars,
e.g., how many cars are insured, the brand and the price of the vehicles, as well
as the record of the claims and the respective compensations; iii)housing, e.g.,
the

number of
private buildings or houses insured, the price of the building and the logs
of the claims. Customers' details are aggregated at
the level of the 170 municipal-
ities for which we have ocial census data using the zip code as spatial reference;
this step leaves us with 568,426 customers matching the geographical boundaries.

Table2summarizes the information available.

- 733 Removals + 810 Additions

Table 3: Features extracted from the insurance database aggregated at the munici-pality level. We extracted features from di↵erent customers information: i) Demo-graphics (age, employment status, etc.), ii) Cars (engine displacement – CCM, price of the vehicle, etc.) and iii) Housing (year of the building, number of claims, etc.)]

CategoryNameDescription
Demographicf1:Unemployment rate
f2:Average age in the municipality
f3:Fraction of owners (house)
f4:Fraction of foreigners
f5:Average number of customers with at least one child
f6:Market share
f7:Fraction of women
f8:Number of customers divided by total customers
Carsf9:Average price of the cars
f10:95th percentile price of the cars
f11:Average year of the car
f12:5th percentile year of the car
f13:Average CCM of the car
f14:95th percentile CCM of the car
f15:Average number of claims per cars
f16:95th percentile number of claims of the car
f17:Average sum of claims of the car
f18:95th percentile number of price of the car
f19:Average premium of the car
f20:Percent of insured cars
Housingf21:Average class of furniture
f22:95th percentile class of furniture
f23:Average number of rooms
f24:95th percentile number of rooms
f25:Average building insured sum
f26:95th building insured sum
f27:Average building year of Construction
f28:5th percentile building year of construction
f29:Average type of building
f30:Average number of claims per building
f31:Average sum of claims per building
f32:95th sum of claims per building
f33:Average Insured Premium
f34:95th sum of insured premium

Table 4: List of the selected indicators for the 6 di↵erent categories.

CategoryTarget Variable

Table 3: Final set of features aggregated at the municipality level.

CategoryNameDescription
Demographicf1:Unemployment rate
f2:Average age in the municipality
f3:Fraction of owners (house)
f4:Fraction of foreigners
f5:Average number of customers with at least one child
f6:Market share
f7:Fraction of women
f8:Number of customers divided by total customers
Carsf9:Average price of the cars
f10:95th percentile price of the cars
f11:Average year of the car
f12:5th percentile year of the car
f13:Average CCM of the car
f14:95th percentile CCM of the car
f15:Average number of claims per cars
f16:95th percentile number of claims of the car
f17:Average sum of claims of the car
f18:95th percentile number of price of the car
f19:Average premium of the car
f20:Percent of insured cars
Housingf21:Average class of furniture
f22:95th percentile class of furniture
f23:Average number of rooms
f24:95th percentile number of rooms
f25:Average building insured sum
f26:95th building insured sum
f27:Average building year of Construction
f28:5th percentile building year of construction
f29:Average type of building
f30:Average number of claims per building
f31:Average sum of claims per building
f32:95th sum of claims per building
f33:Average Insured Premium
f34:95th sum of insured premium

3.2 Swiss Census Data

The ocial statistics for the Swiss municipalities are collected and made available online within the initiativeStatistics of Swiss Citiesthat is is the result of a collabo-ration between the Union of Swiss Cities and the Federal Statistical Oce (FSO)
[3]
.
The report is published in the first quarter of the year and it presents varied facets of the urban life; we focus on six domains:population(p),transportation(t),employ-ment(w),space and territory(s),housing(h) andecon

733 Removals 810 Additions

Populationp
1
:Fractionofforeigners
p
2
:Fractionofbeneficiariesofsocialassistance
Transportationt
1
:Carsper1000inhabitants
t
2
:Fractionofcommutersusingpublictransportation
Employmentw
1
:Unemploymentrate
w
2
:Unemploymentrateamongwomen
Space and Territorys
1
:Areacoveredbybuildings(%)
s
2
:Greenarea(%)
Housingh
1
:Vacancyrate(%)
h
2

housing and 9 for economy. From

the original dataset, we focus on the key target variables that are not redundant and that can be a proxy for quality of life in cities, such as the unemployment rate [24], use of public transportation [39] or investment in culture [40]. As a result of this process, we restrict the analysis to two indicators for each domain: the fraction of foreigners and the rate of beneficiaries of social assistance (p), the number of private cars per person and the fraction of commuters using public transportation (t), the unemployment rate and the unemployment rate among women (w), the percentage of areas covered by buildings or green areas (s), the vacancy rate and the aver-age area per inhabitant (h), and the municipal debt and fraction of investment in culture (e). The complete list of selected target variables is summarized in Table4.

[3] https://www.bfs.admin.ch/bfs/fr/home/statistiques/catalogues-banques-donnees/publications/ouvrages-synthese/statistiques-villes-suisses.html

Donadioetal.Page 7 of30

Table 4: List of the target indicators for the 6 di↵erent domains.

DomainVariable

Populationp
1
:Fractionofforeigners
p
2
:Fractionofbeneficiariesofsocialassistance
Transportationt
1
:Carsper1000inhabitants
t
2
:Fractionofcommutersusingpublictransportation
Employmentw
1
:Unemploymentrate
w
2
:Unemploymentrateamongwomen
Space and Territorys
1
:Areacoveredbybuildings(%)
s
2
:Greenarea(%)
Housingh
1
:Vacancyrate(%)
h
2

- 733 Removals + 810 Additions

2

:Fractionofinvestmentinculture

measured overnspatial units and is given by:

I=
n
s
0
X
i
X
j
z
i
w
i,j
z

2

:Fractionofinvestmentinculture

3.3 Validation

As a validation step, we test the representativeness of the insurance data along four dimensions: (a) total population, (b) percentage of foreigners, (c) percentage of pop- ulation aged 20-40, and (d) percentage of population aged 0-19. Figure2shows, for each dimension, a scatter plot and the corresponding Pearson's correlation coe- cient computed using the ocial census data and the La Mobili`ere customers base.

A high degree of correlation (⇢=0.91) can be observed for the total population variable, meaning that the insurance dataset mimics quite well the population dis- tribution at the municipality level. Focusing on age, we observe a strong relation for the case of customers in the age range 20-40 (⇢=0.8) while the correlation disappears (⇢=0.05) for customers aged 0-19. This behavior is expected since children and teenagers are not usually the owners of insurance policies on vehicles or houses. Last, we observe a solid relation with the percentage of foreigners (⇢=0.6).

It is worth noting that socioeconomic processes often manifest non-random spatial patterns that make close areas more similar than distant ones. Moreover, spatial e↵ects do not apply only to the case of neighboring areas; on the contrary, a consis- tent body of literature in geography define spatial relationships between aerial units as a function of distance [41]. Often this choice depends on prior knowledge about the area under study or a conceptualization of the interactions between neighbor- ing locations with regards to quantity under study. In this work, we refer to the Moran's I statistic [42] to assess the presence of spatial autocorrelation in the census variables. Moran's I measures the global spatial autocorrelation of an attributey measured overnspatial units using the following relation:

I=
n
s
0
X
i
X
j
z
i
w
i,j
z

- 733 Removals + 810 Additions

i
z
j
(1)

Massaroetal.Page 8 of29
A
A
B
CD
Figure 2: Pearson’s correlation between aggregate
d information of the insurance cus-
tomers (x-axis) and the census data (y-axis). Eac
h point corresponds to a di↵erent
municipality. On the top row (A–B) we report the
 correlation between total number
of inhabitants/customers (a) and fraction of insu
red foreigners versus fraction of
foreigners citizens (B). On the bottom (C–D) We s
how the correlation between the
fraction of customers aged between 20–40 years ol
d (C) and 0–19 years old (D) with
the fraction of the population with the same age.
where wherew
i,j
is a spatial weight,z
i
=y
i
¯y, ands
0
=
P
i
P
j
w
i,j
. Spatial
weights are computed using endogenous adaptive ba
ndwidths with a Gaussian ker-
nel function
implemented in the Python packagepysal

i
z
j
(1)
wherew
i,j
are the spatial weights,z
i
=y
i
¯ywith ¯ybeing the average across spatial
units, ands
0
=
P
i
P
j
w
i,j
.
In our experimental settings, the spatial weights
are
computed using endogenous adaptive bandwidths wit
h a Gaussian kernel function

Donadioetal.Page 8 of30
A
A
B
CD
Figure 2: Pearson’s correlation between aggregat
ed information of the insurance
customers (x-axis) and the census data (y-axis).
 Each point corresponds to a di↵er-
ent municipality. On the top, we report the corre
lation between the total number
of inhabitants and customers (A) and between the
 fraction of foreigners versus the
fraction of insured foreigners (B). On the botto

733 Removals 810 Additions

450 [4]
451 . Table5shows how all
452 the selected
target variables are positively spatially autocorrelated, ranging
from
453 I=0.56 toI=0.8.
454 4Methods
455 In our analysis,
we adopt multivariate linear regression to predict
the socio-economic
456 indicators of
interest. We present two spatially aware models that we introduce to
capture the geographical dependencies emerging in
457 our problem.
458 [4]
459 https://pysal.readthedocs.io/en/latest/
460
461 Massaroetal.Page 9 of29
462 Table 5: Moran’s I coecients for the main target variables.
463 VariableMoran’s I
464 p
465 1
466 :Fractionofforeigners0.7
467 p
468 2
469 :Fractionofbeneficiariesofthesocialassistance0.73
470 t
471 1
472 :Carsper1000inhabitants0.56
473 t
474 2
475 :Fractionofcommutersusingpublictransportation0.76
476 w
477 1
478 :Unemploymentrate0.8
479 w
480 2
481 :Unemploymentratebetweenwomen0.79
482 s
483 1
484 :Buildingarea(%)0.74
485 s
486 2
487 :Greenarea(%)0.64
488 h
489 1
490 :Vacancyrate(%)0.66
491 h

c range.
446 implemented in the Python packagepysal
447 [4]
448 . Table5shows how all the selected
target variables are positively spatially autocorrelated, ranging
449 fromI=0.56 to
I=0.8. This implies that municipalities that are
450 closer in space tend to share
451 similar socioeconomic characteristics.
452 4Methods
453 In
this section, we describe the methodological steps of our predictive pipeline. After
454 a features selection module,
we adopt multivariate linear regression to predict
the
455 socioeconomic indicators of
interest using two spatially-aware models that capture
456 the global and local geographical dependencies.
457 [4]
458 https://pysal.readthedocs.io/en/latest/
459
460 Donadioetal.Page 9 of30
461 Table 5: Moran’s I coecients for the main census variables.
462 VariableMoran’s I
463 p
464 1
465 :Fractionofforeigners0.7
466 p
467 2
468 :Fractionofbeneficiariesofthesocialassistance0.73
469 t
470 1
471 :Carsper1000inhabitants0.56
472 t
473 2
474 :Fractionofcommutersusingpublictransportation0.76
475 w
476 1
477 :Unemploymentrate0.8
478 w
479 2
480 :Unemploymentratebetweenwomen0.79
481 s
482 1
483 :Buildingarea(%)0.74
484 s
485 2
486 :Greenarea(%)0.64
487 h
488 1
489 :Vacancyrate(%)0.66
490 h

- 733 Removals + 810 Additions

e
2
:Fractionofinvestmentinculture0.7

4.1 Feature selection
The first step in constructing cost-e↵ective pred
ictors is to select the features that
will best predict a given
variable. For each of the
variables in Table4,asubset
of explanatory variables was selected from all th
e variables present in Table3by
the means of di↵erent selection algorithms. Once
 the features are selected for each

variable, the spatial dimension of the problem wi
ll be integrated to the model using
two di↵erent approaches, as described in subsecti
ons4.2and4.3. The following
three di↵erent feature selections approaches were
explored:
Simple [41]:standard OLS multivariate regression
 model where features are se-
lected according to theirp
value
with a threshold ofp
-
value
=0.05.
Lasso [42]:to reduce model complexity and to prev
ent overfitting we turn to the
Least Absolute Shrinkage and Selection Operator
 (LASSO) model that is well-suited
for cases showing high levels of multicollinearit
y. In particular, we use theLasso-
LarsICclass available in scikit-learn
[5]

that relies on Least Angle Regression and

the Bayes Information Criterion for model
selection and to find a trade-o↵between
the goodness of fit
 and the complexity of the model.
Interactions [43]:this variant is also based on L
asso but it considers the interac-

tions among the features by ensuring that the num
ber of predictorsn

p

are between

defined boundariesb1⇥n

p

⇥b2, whereb1 = 2 andb2 = 15. This is because we
need to ensure in the interactions step that the

e
2
:Fractionofinvestmentinculture0.7

4.1 Features selection
The first step in constructing cost-e↵ective pred
ictors is to select the features that
will best predict a given outcome
variable. For each of the socioeconomic indica-

tors in Table4, we select a subset of explanatory
variables from the initial pool
of covariates summarized in Table3using theLassoL
arsICmodule
 available in

scikit-learn
[5]
. To reduce model complexity and to prevent overf
itting,LassoLarsIC
adopts the Least Absolute Shrinkage and Selection
Operator [43] (LASSO) model

for fit and it relies on theLeast Angle
Regression[44] (LARS) and theBayes Infor-

mation Criterion[45] (BIC) for model
selection, trying to find the right trade-o↵

between fitting performance
 and the complexity of the model.
Since variable selection methods may su↵er from m
odel instability or potential
bias in parameter estimates and confidence interv
als (especially relevant in explana-
tory modeling), we implemented the methodology an
d practical suggestions pro-
posed in [46,47] to control for these e↵ects. In
 particular, we aim at estimating the
stability of the selection procedure to random pe
rturbations of training samples. We
implemented a subsampling without replacement rou

b2
b1

<= 170.

4.2 Spatial Lag Model

In a Spatial Lag Model (SLM), the first and most straightforward way to introduce space is by “spatially lagging” the dependent variable. One must then treat it as

an endogenous variable, this is known as a Spatial Autoregressive Model. Formally.

This can be expressed in matrix notation, as follows [44]:

y=↵+X+W y+✏(2)

whereyis the vector of observations on the dependent variable,Xis the matrix of observations on the exogenous variables,Wis the spatial weighting matrix of known constants,is the vector of regression parameters andis the scalar autoregressive

[5]

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLarsIC.html

Massaroetal.Page 10 of29

parameter. The variableWyis typically known as the spatial lag ofy. The weighting scheme will determine how the spatial dimension of the problem is incorporated in the model. A k-nearest neighbors (KNN) scheme was chosen, and the number of

nearest neighbors to include in the scheme was optimized by maximizing the average

R

2

over all the dependent variables and found to be equal to ten neighbors.

even in cases where the classical bootstrap fails [48]. We performed 500 subsampling iterations and we computed the stability estimator proposed by Nogueira et al. [49] that is a frequency-based statistics, ranging 0 to 1 and monotonically increasing as

the stability of the feature selection grows. The idea is that the stability measure is a linear function of the sample variances with a strictly negative slope. According to the proposed framework, stability values above 0.75 represent an excellent agreement of the feature sets beyond chance, between 0.75 and 0.4 intermediate to good agreement, while values below 0.4 represent a poor agreement.

4.2 Spatial Lag Model

To characterize the influence of neighboring spatial units, we implement a Spatial Lag Model [50] (SLM) where the local e↵ects are encoded adding a term that contains a spatially lagged version of the dependent variable. SLM is an instance of

Spatial Autoregressive Models where the additional term

is treated as an endogenous variable. More formally, this can be expressed in matrix notation, as follows:

[5]

https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoLarsIC.html

Donadioetal.Page 10 of30

y=↵+X+W y+✏(2)

whereyis the vector of observations on the dependent variable,Xis the matrix of observations on the exogenous variables,Wis the spatial weighting matrix of known constants,is the vector of regression parameters andis the scalar autoregressive

parameter. The variableWyis typically known as the spatial lag ofy. The weighting scheme will determine how the spatial dimension of the problem is incorporated in the model. We adopt a

k-nearest neighbors (KNN) scheme

and we optimize the

hyperparameterkwith a grid search approach that aims at finding the optimal

value that maximizes the averageR

2

across all the dependent variables. We use the

733 Removals 810 Additions

coecients the PySAL package was used [45].
4.3 Geographical Weighted Regression
Geographical Weighted Regression (GWR) is a local form of linear regression used
to model spatially varying relationships. In regression analysis, we try to explain
the variations of a dependent variable using a su

ite of uncorrelated and normally
distributed independent variables. The strength and direction of association is in–
dicated by the regression coecients, with one coecient given for each variable in
the dataset. In GWR, instead of one global coecient for each variable, coecients
are able to vary according to space. This spatial variation in coecients can reveal
interesting patterns which otherwise would be masked. The general formula for a
GWR is an extension of Equation3where one regression is calculated for each point
using spatial weights.
y
i
=
0i
+
m
X
j=1

i,j
X
i,j
+ε
i
.(3)
The indexiindicates the location of the city of interest. GWR basically fits a set
ofcoecients for each location:

i
=(X
T
W
i
X)
1
X
T
W
i
y,(4)
whereW
i
is the diagonal matrix of the spatial weights, unique to locationi.There
are various schemes for calculating the weights, nearest neighbors, cubic or expo–
nential kernels. The weights are computed by the following:

4.3 Geographical Weighted Regression
Geographical Weighted Regression (GWR) is a local form of linear regression used
to model spatially varying relationships. In regression analysis, we try to explain
the variations of a dependent variable using a su

ite of uncorrelated and normally
distributed independent variables. The strength and direction of association is in–
dicated by the regression coecients, with one coecient given for each variable in
the dataset. In GWR, instead of one global coecient for each variable, coecients
are able to vary according to space. This spatial variation in coecients can reveal
interesting patterns which otherwise would be masked. The general formula for a
GWR is an extension of Equation3where one regression is calculated for each point
using spatial weights.
y
i
=
0i
+
m
X
j=1

i,j
X
i,j
+ε
i
.(3)
The indexiindicates the location of the city of interest. GWR basically fits a set
ofcoecients for each location:

i
=(X
T
W
i
X)
1
X
T
W
i
y,(4)
whereW
i
is the diagonal matrix of the spatial weights, unique to locationi.There
are various schemes for calculating the weights, nearest neighbors, cubic or expo–
nential kernels. The weights are computed by the following:

611 i,j
612 b
613),(5)
614 whered
615 ij
616 is the Euclidean distance between municipalitiesi andjandbis the
617 bandwidth of the kernel that has to be chosen. For each municipality we calculate
618 a vector of weights and then regress using the final formula (whether linear or
619 with interactions), in order to estimate all indicators for each municipality. The
620 parameters forming the GWR will be the focus of some analysis because of the non-stationarity of the problem. It is interesting to
621 explore how the influence of certain explanatory variables changes from city to city a
622 nd whether there are underlying tendencies. We estimated a fixed bandwith for eve
623 ry city and every model. The optimum bandwidth can be estimated by minimizing
624 an information criterion; in practice, a corrected version of the AIC is used,
625 which unlike basic AIC is a function
626
627 Massaroetal.Page 11 of29
628 of sample size [46]. Thus for a GWR model with a bandwidthb,theAIC
629 c
630 is given
631 by:
632 AIC
633 c
634 =2nln ˆ+nln 2↑+n
635 n+tr(S)
636 n2tr(S)
637 ,(6)
638 where ˆis the estimated standard deviation of the residuals;nis the number of
639 observations andtr(S) is the trace of the hat matrixS. The hat matrix is the
640 projection matrix from the observedyto the fitted

600 i,j
601 b
602),(5)
603 whered
604 ij
605 is the Euclidean distance between municipalitiesi andjandbis the
606 bandwidth of the kernel that has to be chosen. For each municipality we calculate a vector of weights and then regress using the fina
607 l formula (whether linear or with in-
608 teractions), in order to estimate all indicators for each municipality. The parameters
609 forming the GWR will be the focus of the analysis because of the non-stationarity
610
611 Donadioetal.Page 11 of30
612 of the problem. It is interesting to explore how the influence of certain explanatory
613 variables changes from city to city and whether there are underlying tendencies. We
614 estimate a bandwidth for each city and model. The optimal bandwidth is estimated
615 by minimizing an information criterion; in practice, we adopt a corrected version
616 of the AIC that, in contrast with the original definition, is a function of sample
617 size [51]. In more details, in a GWR model with a bandwidthb,theAIC
618 c
619 is given
620 by:
621 AIC
622 c
623 =2nln ˆ+nln 2↑+n
624 n+tr(S)
625 n2tr(S)
626 ,(6)
627 where ˆis the estimated standard deviation of the residuals;nis the number of
628 observations andtr(S) is the trace of the hat matrixS. The hat matrix is the
629 projection matrix from the observedyto the fitted

- 733 Removals + 810 Additions

=X
i
(X
T
W
i
X)
1
X

T
W
i
(7)

4.4 Cross Validation

The regression parameters and hyper-parameter selection is performed on the train-ing set (80%) using cross-validation and then once the model is fitted, the final performance is tested on the remaining 20% of the data. The municipalities were partitioned into three classes: communes with less than 25k inhabitants,those with between 25 and 100k, and those with a population higher than 100k. Note that there are five cities that belong to the last class: Z¨urich, Gen`eve, Basel, Lausanne and Bern, which represent the five main Swiss cities. In each round of the cross-validation, the procedure ensures that each fold is a fair representation of the whole distribution. A training set with 80% of the points and a validation set with the remaining 20% is selected. The training set is used to calibrate the parameters for the regression and the hyper-parameters of the GWR (bandwidth selection) and SLM (spatial lags). The validation set is used to test the ability of the model to be generalized to unknown locations. We repeat this procedure five times by ensuring that the municipalities in the third class appear only once for each training phase and that the same municipalities don't appear in both the training and i the valida-

tion phase. In the results section we report the average and the standard deviation of the models performances due to cross-validation.

5Results

In the first part of the section, we show the results of the features selection process. In particular, in Table6we report the significant features for each target variable

using the Lasso method. The number of selected features varies between 2 (for the

housing variableh

=X
i
(X
T
W
i
X)
1
X

T
W
i
(7)

4.4 Evaluation

To test the performance of the predictive pipeline we refer to an out-of-sample vali-dation where the estimation of the regression parameters and the hyper-parameter tuning are performed on a training set using a 80% split and cross-validation, while the predictive performance is tested on the hold-out data (remaining 20%). To cope with the heterogeneity of the population distribution in our sample and to allow to train and test the models with a sample that is representative of the entire spec-trum of population size, we implement a stratification approach. It is worth noting that using a random sampling strategy instead, we could end up in the cross valida-tion procedure with splits that contain only highly or low populated municipalities, introducing a bias in the evaluation pipeline. In this direction, we partition the mu-nicipalities in three classes: communes with less than 25k inhabitants, those between 25k and 100k, and those with a population higher than 100k. Note that there are five cities that belong to the last class: Z¨urich, Gen`eve, Basel, Lausanne and Bern, which represent the five main Swiss cities. In each round of the cross-validation, the procedure ensures that each fold is a fair representation of the whole distri-bution balancing the three classes. We adopt a 5-cross validation accordingly. In

the results section, we report the average and the standard deviation of the models

performance due to cross-validation.

5Results

In the first part of this section, we present the results of the features selection process for each of the target indicators. After applying theLassoLarsICmethod

the number of selected features spans from 2 (for the Housing indicatorh

2
). In the Population cate-
gory, the per percentage of foreigners (p

we present the selected features for each model.
 Overall, we observe a fair degree
of robustness to random perturbations with the me
asure of stability that varies
across dimensions. In particular,p
1
(0.82) andt
2
(0.77) show the highest stability
that reaches an excellent level of agreement;w

2
,h
2
,w
1
,p
2
,e
2
,s
2
,t
1
,s
1
, andh
1

Donadioetal.Page 12 of30
Table 6: Summary of the results of the Spatial La
g Model (SLM) for the target
indicators in the domains Population, Transportat
ion, and Employment.
VariableFeaturesCoecient Probability [0.025
0.975]
p1Intercept−0.0039.48E−01 −0.103 0.097
Fraction of foreignersf3−0.1191.24E−01 −0.271
0.034
f40.4601.65E−12 0.332 0.589
f70.2651.06E−04 0.130 0.400
f23−0.1091.74E−01 −0.267 0.049
Wdepvar0.0194.24E−01 −0.028 0.067
p2Intercept−0.0098.49E−01 −0.101 0.083
Fraction off10.2981.23E−07 0.187 0.409
beneficiaries off9−0.4171.39E−02 −0.752 −0.08
2
social assistancef130.1344.45E−01 −0.213 0.4
82
f150.1571.51E−02 0.029 0.284
f23−0.3276.13E−10 −0.431 −0.223
Wdepvar0.0321.49E−01 −0.012 0.076
t1Intercept0.0088.84E−01 −0.101 0.117
Cars perf1−0.1372.94E−02 −0.260 −0.013
1000 inhabitantsf60.3811.48E−09 0.256 0.505
f7−0.1191.08E−01 −0.264 0.027
f8−0.1304.98E−02 −0.261 0.001
f170.0762.95E−01 −0.068 0.221
f190.1651.69E−02 0.029 0.301
f200.0683.69E−01 −0.082 0.217
f21−0.1384.38E−02 −0.272 −0.003

-0.138
using publicf6-0.1011.63E-02 -0.184 -0.018
transportationf20-0.2611.39E-07 -0.358 -0.163
f220.1201.39E-02 0.024 0.217
f250.0913.64E-02 0.005 0.176
Wdepvar0.1193.23E-16 0.090 0.148
w1Intercept-0.0088.62E-01 -0.097 0.082
Unemployment ratef10.2201.26E-04 0.106 0.333
f40.1876.07E-04 0.079 0.295
f70.2061.18E-03 0.081 0.332
f13-0.1251.25E-02 -0.225 -0.026
f23-0.2121.36E-04 -0.322 -0.102
Wdepvar0.0541.26E-02 0.011 0.096
w2Intercept-0.0127.74E-01 -0.096 0.071
Unemployment ratef10.1514.60E-03 0.046 0.255
between womenf40.1503.05E-03 0.050 0.251
f6-0.0876.45E-02 -0.181 0.006
f70.1067.38E-02 -0.011 0.224
f9-0.2501.39E-04 -0.379 -0.120
f160.0335.13E-01 -0.066 0.132
f190.2866.83E-05 0.144 0.428
f23-0.1245.06E-02 -0.250 0.001
f33-0.2685.80E-06 -0.384 -0.151
Wdepvar0.0612.60E-03 0.021 0.101
cover a range between good (0.72) and intermediate (0.43) stability (variables are
listed in decreasing order), whilee
1
(0.22) shows a poor agreement. This low value
indicates how the model characterizing the municipal debte
1
is highly dependent
on variations of the training set to define significant determinants. Consistently,e
1
is also the indicator with the lowest performance in the predictive task, indicating
how the insurance data is not really able to capture its behavior.
Switching the focus on the predictive task, Table 6and Table7summarize the
results of the Spatial Lag Model for all the indicators. We present the direction and

Donadioetal.Page 13 of30
Table 7: Summary of the results of the Spatial Lag Model (SLM) for the target
indicators in the domains Space and Territory, Housing, and Economy.
VariableFeaturesCoecient Probability [0.025 0.975]
s1Intercept-0.0325.94E-01 -0.150 0.086
Building area (%)f3-0.2862.29E-04 -0.439 -0.133
f40.0752.88E-01 -0.064 0.213
f70.1384.31E-02 0.003 0.274
f250.1001.16E-01 -0.026 0.226
f310.1347.74E-02 -0.016 0.283
Wdepvar0.0702.53E-02 0.008 0.132

- 733 Removals + 810 Additions

1
) is strongly related to the featuref
4
that
characterizes the fraction of foreign customers
of La Mobili`ere as expected. This
provides an additional validation on the
representativeness of the dataset used in
this analysis.
Moreover, the percentage of people that receive
social assistance (p
2
)
is strongly linked to the unemployment ratef
1
, providing another reasonable ex-
planation, and to the average number of roomsf

Wdepvar0.0886.98E-05 0.044 0.132
h1Intercept0.0395.54E-01 −0.091 0.169
Vacancy rate (%)f30.0436.29E-01 −0.132 0.217
f200.1672.54E-02 0.019 0.314
Wdepvar0.1571.19E-04 0.076 0.237
h2Intercept0.0167.03E-01 −0.067 0.099
Average areaf1-0.1611.66E-03 −0.263 −0.060
per inhabitantf20.1425.85E-03 0.040 0.244
in square metersf30.1748.51E-03 0.043 0.305
f4-0.0305.68E-01 −0.133 0.073
f60.1181.06E-02 0.027 0.210
f210.1372.44E-02 0.017 0.256
f220.0523.70E-01 −0.062 0.165
f230.3133.30E-06 0.180 0.446
f27-0.0821.07E-01 −0.182 0.018
Wdepvar0.0631.14E-04 0.031 0.096
e1Intercept-0.0148.31E-01 −0.143 0.115
Municipal debtf9-0.1062.04E-01 −0.270 0.059
f160.0981.84E-01 −0.047 0.243
f270.2364.07E-04 0.104 0.368
Wdepvar0.0811.03E-01 −0.017 0.180
e2Intercept-0.0128.23E-01 −0.114 0.091
Fraction of investmentf10.1229.03E-02 −0.020
0.265
in culturef40.0029.72E-01 −0.138 0.143
f70.0009.98E-01 −0.144 0.144
f11-0.2091.59E-02 −0.380 −0.038
f120.0297.49E-01 −0.148 0.205
f21-0.1001.20E-01 −0.228 0.027
f23-0.1287.58E-02 −0.271 0.014
f260.1079.72E-02 −0.020 0.234
f340.0366.35E-01 −0.113 0.184
Wdepvar0.1318.81E-09 0.086 0.176
the intensity of the relations along with confide
nce intervals; significant determi-
nants are marked in bold.
In the Population domain, the fraction
of foreigners (p
1
) is positively (=0.46)
linked to the demographic featuref
4
that represents the fraction of foreigners cus-
tomers of La Mobili`ere
and, in the same direction, to the fraction of
womenf
7
(=0.265).
Moreover, the percentage of people that receive
social assistance (p
2
)
is positively linked to the unemployment ratef
1
(=0.298), and the average num-
ber of claims per carf
15
(=0.157). We observe a negative relation with ave
rage

733 Removals 810 Additions

that is indirectly connected to
the social class. The most relevant features in t
he Transportation category are not
of immediate interpretation. In fact, for the var
iablet
1
(cars per 1000 inhabitants),
the most relevant feature is the market share (f
6
) while for the variablet
2
(com-
muters that use public transportation) the most r
elevant feature is the fraction of

Massaroetal.Page 12 of29
Table 6: Results of the Lasso features selection
 for the di↵erent target variables.
The features a are order in an ascending of the p
-value (lowest to highest).
CategoryVariable   Feature   p-valueVariable   Fe
ature   p-value

Populationp

1
f4

⇤⇤⇤
1.02E-14p
2
f23
⇤⇤⇤
2.01E-09
f7

⇤⇤⇤
4.07E-06f1
⇤⇤⇤
9.94E-09
f31.07E-01f15
⇤⇤
3.40E-03
f232.32E-01f9
⇤

(=0.327)

as indirect proxy for the social class.
In the Transportation domain, the number of cars
 per 1000 inhabitantst
1
shows
a negative relation with the unemployment ratef
1
(=0.137) and the average
class of furnituref

21
(=0.138). A positive link is found with the marke
t share

Donadioetal.Page 14 of30
f
6
(=0.381), the average premium of the carsf
19
(=0.165) and the average
years of constructionf
27
(=0.154). Concerning the commuters that use publi
c
transportationt
2
we observe a negative link with the market sharef

6
(=0.101)
and the fraction of house ownersf

3

(=0.239). This could be explained by
the observation that individuals living in rental
houses show a higher frequency of
ride-sharing use and commuting using public trans
portation than those who own
their houses [53]. A positive relation is found w
ith the percent of insured carsf
20
(=0.261), the 95th percentile of the class of the
insured furnituref
2
2(=0.120),
and the average insured sum per buildingf
25
(=0.091).
Focusing on the Work category, the unemployment r
atew

733 Removals 810 Additions

```
⇤⇤⇤
2.33E-08t
2
f3
⇤⇤⇤
1.69E-07
f1
⇤
2.19E-02f22
⇤⇤⇤
9.42E-07
f19
⇤
2.41E-02f20
⇤⇤⇤
3.55E-06
f87.93E-02f6
⇤
4.35E-03
f72.85E-01f25
⇤
1.84E-02
f179.63E-02
f203.74E-01
Workw
1
f1
⇤⇤⇤
6.77E-06w
2
f19
⇤⇤⇤
6.03E-08
f7
⇤⇤⇤
1.78E-05f9
⇤⇤⇤
4.34E-06
f4
⇤⇤⇤
2.59E-05f33
⇤⇤⇤
1.39E-05
f23
⇤⇤
7.80E-04f1
⇤⇤
5.02E-04
f13
⇤
1.11E-02f4
```

```
is positively connected
to a set of demographics features, primarily the
 fraction of foreignersf
4
(=0.187),
the fraction of womenf
7
(=0.206) and, as expected, the unemployment rate
 of
the Mobili`ere customersf
1
(=0.220). We observe an opposite relation with th
e
average CCM of the carsf
13
(=0.125), and the average number of roomsf
23
(=0.212). For the case of the women unemployment
 ratew
2
, the dominant
features are related to the economic characterist
ics of the items insured, being
the average premium of the carsf
19
(=0.286) in a positive relation and the
average price of the carsf
9
(=0.250) or average insured premiumf
33
(=
0.268) linked negatively. These observations tend
to indicate gender di↵erences in
the insurance sector. The fraction of foreignersf
4
(=0.150) and the customers
unemployment ratef
```

- 733 Removals + 810 Additions

f7
⇤
3.04E-02
f236.13E-02
f164.07E-01
Space and Territorys

1

f3
⇤⇤⇤
1.22E-05s

(=0.151) behave accordingly to the previous case. Within the Space and Territory category, both the variables percentage of building
areas
1
and percentage of green areas

2

f3
⇤⇤⇤
2.89E-05
f31
⇤
2.85E-03f11
⇤
1.66E-03
f7
⇤
1.55E-02f7
⇤
1.49E-02
f4
⇤
4.75E-02f233.68E-01
f251.36E-01
Housingh

are negatively connected to the fraction of
house ownerf
3
(=0.286 and=0.270 respectively).
In the Housing domain, the vacancy rateh

1

f3
⇤⇤
7.67E-03h

appears to be positively related to the
percentage of insured carsf
20
(=0.167), while the average area for inhabitant
h

2

f23
⇤⇤⇤
2.34E-06
f7
⇤
1.55E-02f1
⇤⇤⇤
1.80E-05
f4
⇤
4.75E-02f3
⇤
3.94E-03
f20
⇤
8.25E-02f27
⇤
4.61E-03
f251.36E-01f6
⇤
1.44E-03

show several positive links to the average agef

- 733 Removals + 810 Additions

1.54E-02
f4
*
2.36E-01
f227.89E-01
Economye
1
f27
**
5.47E-04e
2
f1
***
1.17E-06
f9
*
1.05E-02f21
*
6.10E-03
f16
*
4.07E-02f34
*
1.28E-02
f4
*
2.48E-02
f11
*
3.49E-02
f236.60E-02
f71.46E-01
f252.12E-01
f129.72E-01
***<0.0001,**<0.001and *<0.05.
customers that own an house (f
3
). For a possible explanation in this direction, it
has been observed that individuals living in rental houses show a higher frequency
of ridesharing use and commuting using public transportation than those who own
their houses [48]. In the Work category, the unemployment ratew
1
is primarily
connected to a set of demographics features, e.g.,f
6
andf
1

2
(=0.142), the fraction of
house ownersf
3
(=0.174), the market sharef
6
(=0.118), the average class
of furnituref
21
(=0.137) and the average number of roomsf
23
)(=0.313).
Higher values forh
2
corresponds to lower unemployment ratef
1

904 2
905 , the domi-
906 nant features are related to the
economic characteristics of the
objects insured, e.g.,
907 f
908 19
909 average sum of class premium of the car orf
910 9
911 average price of the cars orf
912 3
913 3
average insured premium, that tend to indicate ge
914 nder di↵erences in the insurance
sector. Within the Space and Territory category b
915 oth variabless
916 1
917 building area
918 ands
919 2
green area are strongly connected to the fraction
920 of house ownerf
921 3
922 that
suggests a link between urban characteristics of
923 neighborhoods and their average
population. A similar observation applies to the
924 Housing category especially for the
925
926 Massaroetal.Page 13 of29
Figure 3: (A–B). Comparison between the spatial r
927 egression models (bars) and stan-
dard multivariate linear regression (lines) for t
928 he di↵erent features selection models.
(A) Spatial Lag Model and (B) Geographical Weight
929 ed Regression. (C–D) Perfor-
mance using stratified cross-validation for the f
930 ull (black triangles), the training
and the validation sets respectively. (C) Spatial
931 Lag Model and (D) Geographical
932 Weighted Regression.
933 variableh
934 1
vacancy rate. Finally, in the Economy category th
935 e municipal debte
936 1

950 is positively related to the
951 average year of constructions of the buildingsf
952 27
953 (=0.236) that is in accordance
with the literature where modern buildings have b
954 een considered a proxy for eco-
nomic status [54]. Moreover, the fraction of inve
955 stment in culturee
956 2
957 is negatively
958 connected to the average year of the carf
959 1
960 1(=0.209).
It is worth noting that for a group of indicator
961 s, the corresponding predictive
models identify significant relations with expect
962 ed determinants: this is the case of
963 the pair (p
964 1
965 ,f
966 4
) where the fraction of foreigners is explained u
967 sing the information
on the nationality of La Mobili`ere customers. A
968 similar case happen for the pairs
969 (p
970 2
971 ,f
972 1
973), (w
974 1

the municipality that is in accordance with the l
iterature where modern buildings
have been considered a proxy for the economic sta
tus of a city [49]. The fraction
of investment in culturee
2
is connected to the unemployment rate (f
1
) and other
measures of wealth such as the average class of f
urnituref
21
and the sum of the
insured premiumf
34
.
After the first phase of variables selection, we
 compare
 the performance of the
spatially-aware models with a standard multivaria
te linear regressor. Performance
is measured using the coecient of determination p
seudoR
2
for both the space-
agnostic and GWR
models. As shown in Figure3(A-B), both the geogra
phical
models outperform OLS across features selections
 methods and target variables,
with a gain in performance up to 30%. It is worth
noting that GWR is able to
achieve satisfactory results across categories wi
th values ranging from 0.49 fors

2
,f
1
). However, we think that these not surprising re
lations
do not undermine the validity of the experimental
settings for several reasons. First,
the considered models identify alternative predic
tors that are complementary and
cross-domain to the target indicators. Second, th
e observation that a variable con-
structed from a sample of customers of an insuran
ce company is able to predict a
census indicator at the national level is not tri
vial. This represents another sugges-
tion of the validity of
the data collected as a proxy for socioeconomic s
tatus. Third,

Donadioetal.Page 15 of30
A
BC
BC
Figure 3: (A). Comparison of the performance betw
een the spatial regression models
(GWR and SLM) and standard multivariate linear re
gression (OLS). (B-C) Perfor-
mance using stratified cross-validation for the f
ull (black triangles), the training
and the validation sets respectively. (B) Geograp
hical Weighted Regression and (C)
Spatial Lag Model.
to quantitatively evaluate the impact of these no
t surprising variables, we compare
the performance of the original models with a var
iation where we remove them. The
accuracy in terms ofR
2
for both the SLM and GWR
models remains substantially
stable for all the indicators, with an average pe
nalty of 0.034 and 0.014 for SLM
and GWR, respectively. Refer to FigureA.13for a d
etailed comparison.
After the analysis of determinants, we focus on c
omparing the performance of the
global (SLM) and local (GWR) spatial models to a
 standard multivariate linear
regressor (OLS) to quantify to benefit of exploit

963 to 0.83 in the case ofw
964 1
965 orh
966
2
967 . This provides a hint on the ability of
insurance
968 customers data to characterize socio-economic
processes embedded in space.
969 Allowing the relationships
between the independent and dependent variables
to
970 vary by locality to capture contextual
factors, GWR is useful as an exploratory
971 technique; however, its usefulness
as a prediction tool is debated when it comes to
972 model generalizability. To test the
ability to perform out-of-sample predictions we

973 turn to
cross-validation. We focus on the Lasso features
selection case that provides
in general the best performance across categories
974 (see Figure3(A-B)). Since popu-

lation varies broadly across municipalities, we h
975 ypothesize that training and testing

on instances of cities of very di↵erent size coul
976 d introduce a deterioration of perfor-
977 mance. To cope with this e↵ect, we implement a
stratified 5 fold cross-validation in
which the stratification process is regulated by
978 population size. Figure3(C-D) shows
the results of the regression task using stratifi
979 ed cross-validation. As expected, the
980 general performance deteriorates;
however, especially for certain target
variables,
981 we are still able to achieve a good
performance on the validation set, e.g.,h
982 2
983 =0.6,
984 w
985 1
986 =0.53,p
987 2
988 =0.49, andt
989 2
990 =0.49.

1010 . As shown in Figure3(A),
both the spatial models outperform OLS across tar
1011 get indicators, with a gain in
performance up to 30%. It is worth noting that GW
1012 R is able to achieve satisfactory
results across categories with values ranging fro
1013 m 0.49 fors
1014 2
1015 to 0.83 in the case
1016 ofw
1017 1
1018 orh
1019
2
1020 . This provides a hint on the potential of
insurance customers data to
1021 characterize socioeconomic processes embedded in
space. Allowing the relationships
1022 between the independent and dependent variables
to vary by locality to capture con-

1023 textual factors, GWR is useful as an exploratory
technique; however, its usefulness
1024 as a prediction tool is debated when it comes to
model generalizability. To test the
1025 ability to perform out-of-sample predictions we
turn to stratified cross-validation as
described in Section4.4. As shown in Figure3(B-
1026 C),
we observe a decrease of the

1027 overall performance;
however, especially for certain target
variables, we are still able

1028 to achieve a reasonable
performance on the validation set, for example,h
1029 2
1030 =0.6,
1031 w
1032 1
1033 =0.53,p
1034 2
1035 =0.49, andt
1036 2
1037 =0.49.
The values of the performances of the models
1038 are also reported in the Appendix in Table??and T

733 Removals 810 Additions

6 Discussion
In the first part of the paper, we showed how to predict a wide range of socioeco-
nomic indicators using insurance customers activity logs. In this section we shift the
attention to a specific use case that has a strong impact on urban mobility and citi-
zens well being: the relation between commuting and public transport (the variable
t
2
in our settings). The use of public transportation is an important contributing
factor to urban sustainability; it has a heavy environmental footprint reducing air
pollution and trac congestion, among the others. It has also positive financial
benefits for families and communities as a whole, higher level of security and direct
positive e↵ects on well-being and healthier habits. We chose transportation to exem-
plify our data analysis as it is the third most i

to a set of baseline models in which each target indicator is predicted using the
remaining variables from the census. For instance, let us model the fraction of
foreignersp
1
using the explanatory variablesp
2
,t
2
,...,e
2
from Table4. In Figure4

we report a comparison between the performance of the census-based baseline and
the insurance-based models for the cases of SML and GWR. We observe overall
Figure 4: Comparison between the census-based and the insurance-based explana-
tory models for the SML and GWR cases. Positive and negative values mean, respec-
tively, an increase or decrease in performance using La Mobili`ere data in comparison
to the census baseline.
a comparable performance using our approach, with the baseline having a positive
delta of 0.019 on average across indicators. This is expected being the baseline
based on ocial census data where cross-correlation e↵ects are present. However, it
is worth nothing that in our reference scenario the census is not available and, as
such, the baseline approach not feasible. The observation that insurance customers
records are able to achieve comparable results is an additional proof of the potential
of this approach.
6 Discussion
In the first part of the paper, we showed how to predict a wide range of socioeco-
nomic indicators using insurance customers activity logs. In this section we shift the
attention to a specific use case that has a strong impact on urban mobility and citi-
zens well being: the relation between commuting and public transport (the variable
t
2
in our settings). The use of public transportation is an important contributing
factor to urban sustainability; it has a heavy environmental footprint reducing air
pollution and trac congestion, among the others. It has also positive financial
benefits for families and communities as a whole, higher level of security and direct
positive e↵ects on well-being and healthier habits. We chose transportation to exem-
plify our data analysis as it is the third most i

As such, the question of which variables are able to predict the use of
1009 public transport is a key issue. In Table7and in Table8we report the values of the predictors for the global and the GWR models
1010 ls
respectively. While for the
1011 global
model, the parameters have the same values for each municipality, for
the
1012 GWR
we reported the average values of the coecient, the standard deviation,
and
1013 minimum
and maximum values. For the GWR, we observe a high variability in the
1014 intercept, this is mainly due to the high spatial autocorrelation. The GWR adapts
1015 the intercept so it is closer to its neighbors, and thus achieves higher accuracy. More
1016 detailed diagnostic information on the regression, such as the kernel bandwidth is
1017 provided in Table9. Turning the attention to the coecients, we observe that the
1018 fraction of customers that own a house (f
1019 3
1020) is negatively correlated with the target
1021 variable: as expected, it has been observed that individuals living in rental houses
1022 show a higher frequency of ride sharing and public transportation adoption than
1023 house owners [48]. As expected, also the percentage of insured cars (f
1024 20
1025) is nega–
1026 tive linked to the probability of commuting via public transport; the more cars an
1027 individual possesses the less she turns on the public system when it comes to mobil–
1028 ity. Moreover, we observed a higher public transport adoption in major cities, e.g.,
1029 Zurich, Basel, Bern and Geneva as shown in Figure4(b). This is consistent with our
1030 analysis, in fact, the fraction of house owners (f

As such, the question of which variables are able to predict the use of
1082 public transport is a key issue. In Table6and in Table8we report the values of the predictors for the global and the GWR models
1083 ls
respectively (an analysis of the GWR statistics for the target variables is reported in the Appendix from Fig–
1084
ureA.1to FigureA.12and in TableA.3and TableA.4). While for the global
1085
1086
1087 Donadioetal.Page 17 of30
model, the parameters have the same values for each municipality, for
1088
the GWR
we reported the average values of the coecient, the standard deviation,
1089
and min–
1090 imum
and maximum values. For the GWR, we observe a high variability in the
1091 intercept, this is mainly due to the high spatial autocorrelation. The GWR adapts
1092 the intercept so it is closer to its neighbors, and thus achieves higher accuracy. More
1093 detailed diagnostic information on the regression, such as the kernel bandwidth is
1094 provided in Table9. Turning the attention to the coecients, we observe that the
1095 fraction of customers that own a house (f
1096 3
1097) is negatively correlated with the target
1098 variable: as expected, it has been observed that individuals living in rental houses
1099 show a higher frequency of ride sharing and public transportation adoption than
1100 house owners [53]. As expected, also the percentage of insured cars (f
1101 20
1102) is nega–
1103 tive linked to the probability of commuting via public transport; the more cars an
1104 individual possesses the less she turns on the public system when it comes to mobil–
1105 ity. Moreover, we observed a higher public transport adoption in major cities, e.g.,
1106 Zurich, Basel, Bern and Geneva as shown in Figure5(b). This is consistent with our
1107 analysis, in fact, the fraction of house owners (f

733 Removals 810 Additions

2
.
VariableCoecientStd.Errorz-StatisticProbability
Intercept0.1960.1151.6930.089
f3: Fraction of owners (house)-0.3290.141-2.3240.020
f6: Market Share-0.3940.118-3.3350.000
f20: Percent of insured cars-0.4270.138-3.0960.001
f22: 95th percentile class of furniture   0.3640.1332.7270.006
f25: Average Building Insured Sum0.2140.1221.7650.078
Table 8: Summary statistics for the GWR parameters for predictingt
2
.
VariableMeanSTDMinMax
Intercept0.0050.313-1.1690.339
f3: Fraction of owners (house)-0.3060.156-0.6490.007
f6: Market Share-0.2360.218-0.8960.202
f20: Percent of insured cars-0.3010.201-0.5860.541

that seems to be significative is the market share (f
6
). We believe that the market
share feature is representative information because even if, the insurance company
la Mobiliere is a national company, is not used equally across the Swiss country be-
cause of the competition between di↵erent insurance companies. Moreover, having
an insurance contract is mandatory also for renting an apartment, the information
of the market share of a given company can tell us important information about
a certain kind of population living in that area. One of the main characteristic of

GWR is that the inferred relationships vary by locality, that implies each munici-
pality has a di↵erent fitting performanceR
2
and coecients. In Figure5we show
spatial distribution of GWR accuracy in di↵erent regions. Mapping the localR
2
values could provide a useful tool to identify areas where the independent variables
might or might not explain the phenomenon under study. This could be useful, for
example, to identify contextual anomalies that are linked to specific characteristics
of a community. While we are able to achieve good results in several cities, the per-
formance for the Grisons and Ticino cantons are low. These cantons are fairly small
and isolated regions. For example, Ticino is highly influenced by the adjacency to
Italy this influence is not captured by the model. Two of the main cities in Ticino;
Lugano and Belinzona have a very low use of public transport as shown in Figure5.
Another interesting aspect is that the local R
2
shows a clustered behavior, with
adjacent areas having similar performance. Note that these clusters tend to match
administrative boundaries and we can clearly distinguish regions such as Lausanne,
Basel and St. Gallen (light blue), central Switzerland (orange) and the Valais (dark
Table 8: Summary statistics for the GWR parameters for predictingt
2
.
VariableMeanSTDMinMax
Intercept0.0050.313-1.1690.339
f3: Fraction of owners (house)-0.3060.156-0.6490.007
f6: Market Share-0.2360.218-0.8960.202
f20: Percent of insured cars-0.3010.201-0.5860.541

- 733 Removals + 810 Additions

1054 Massaroetal.Page 15 of29
1055 A
1056 B
1057 Figure 4: (A) Spatial distribution of the local coecient of determinationR
1058 2
1059 using
1060 GWR to predict the fraction of commuters using public transportationt
1061 2
1062 .(B)
1063 Comparison between predicted and actual values of the percentage of commuters
1064 using public transport.
1065 prices are higher and people tend not to settle and start a family. Another variable
1066 that distinguishes rural from urban environments is the market share (f
1067 6
1068). Since it
1069 is the oldest insurance company in Switzerland, La Mobiliere has reached customers
1070 all across the country; however, in major cities it has a lower market share, due to
1071 the fiercer competition with other companies and the higher incidence of short-term
1072 and foreigner dwellers. One of the main characteristic of GWR is that the inferred
1073 relationships vary by locality, that implies each municipality has a di↵erent fitting
1074 performanceR
1075 2
1076 and coecients. In Figure4we show spatial distribution of GWR
1077 accuracy in di↵erent regions. Mapping the localR
1078 2
1079 values could provide a useful
1080 tool to identify areas where the independent variables might or might not explain
1081 the phenomenon under study. This could be useful, for example, to identify contex-
1082 tual anomalies that are linked to specific characteristics of a community. While we
1083 are able to achieve good results in several cities, the performance for the Grisons
1084 and Ticino cantons are low. These cantons are fairly small and isolated regions. For
1085 example, Ticino is highly influenced by the adjacency to Italy this influence is not
1086 captured by the model. Two of the main cities in Ticino; Lugano and Belinzona have
1087 a very low use of public transport as shown in Figure4. Another interesting aspect
1088 is that the local R
1089 2
1090 shows a clustered behavior, with adjacent areas having similar
1091 performance. Note that these clusters tend to match administrative boundaries and

1151 Donadioetal.Page 18 of30
1152 A
1153 B
1154 Figure 5: (A) Spatial distribution of the local coecient of determinationR
1155 2
1156 using
1157 GWR to predict the fraction of commuters using public transportationt
1158 2
1159 .(B)
1160 Comparison between predicted and actual values of the percentage of commuters
1161 using public transport.

- 733 Removals + 810 Additions

phenomenon might be linked to the inherent divers
ity in the communities leaving
in the di↵erent areas of the country.
Limitations.The approach proposed in this paper h
as few limitations that should
be carefully discussed. In details:
Sample bias. Even if we showed a fair level of re
presentativeness along di↵erent
dimensions, the input dataset contains informatio
n only on the fraction of popu-
lation that owns an insurance policy with a speci
fic company. Several segments of

Massaroetal.Page 16 of29
Table 9: Information on the GWR oft
2
; percentage of commuters using public

transport.
Diagnostic Information
Spatial kernel:Fixed Gaussian
Bandwidth used:29.030
Residual sum of squares:36.026
E↵ective number of parameters (trace(S)): 41.95
2
Degree of freedom (n − trace(S)):128.048
Sigma estimate:0.53
Log-likelihood:-109.337
AIC:304.578
AICc:334.533
BIC:439.268
R2:0.788
Adj. alpha (95%):0.007
Adj. critical t value (95%):2.723
the population are left out of the analysis, addi
ng a validity bias in the results,
especially for indicators that cover a wider spec
trum of the society.
Spatial granularity
mismatch. Ocial statistics are available at the
level of
municipality and only for a subset of the commune
s, while the insurance customers
data provides information at the finer granularit
y of zip codes. From one side, we
have complete knowledge for a subset of the area
s, while from the other side, a
partial view with a higher coverage. Our analysis
is restricted to the intersection

Table 9: Information on the GWR oft
2
; percentage of commuters using public

transport.
Diagnostic Information
Spatial kernel:Fixed Gaussian
Bandwidth used:29.030
Residual sum of squares:36.026
E↵ective number of parameters (trace(S)): 41.95
2
Degree of freedom (n − trace(S)):128.048
Sigma estimate:0.53
Log-likelihood:-109.337
AIC:304.578
AICc:334.533
BIC:439.268
R2:0.788
Adj. alpha (95%):0.007
Adj. critical t value (95%):2.723
blue) in Figure5. This phenomenon might be linked
to the inherent diversity in the
communities leaving in the di↵erent areas of the
country.
Limitations.The approach proposed in this paper h
as few limitations that should
be carefully discussed. In details:
Sample bias. Even if we showed a fair level of re
presentativeness along di↵erent
dimensions, the input dataset contains informatio
n only on the fraction of popu-
lation that owns an insurance policy with a speci
fic company. Several segments of
the population are left out of the analysis, addi
ng a validity bias in the results,
especially for indicators that cover a wider spec
trum of the society.
Spatial granularity
mismatch. Ocial statistics are available at the
level of
municipality and only for a subset of the commune
s, while the insurance customers
data provides information at the finer granularit
y of zip codes. From one side, we
have complete knowledge for a subset of the area
s, while from the other side, a
partial view with a higher coverage. Our analysis
is restricted to the intersection

1130 heterogeneity of social processes at a micro-leve
l, e.g., neighborhoods in cities.
1131 Temporal evolution. In our analysis we currently
focus on a static snapshot cov-
1132 ering a year of statistics. However, socioeconomi
c conditions vary over time and in
1133 which extent and how fast this change is reflecte
d into the insurance data records
1134 is something not explored yet.
1135 Data availability and privacy. The current approa
ch is based on the assumption
1136 that customers data is available to the researche
rs to tackle relevant challenges that
1137 have a broad social impact. This raises two main
issues related to privacy and the

1138 compliance to the current legislation especially
in the European framework, and
1139 the sharing policy. Proprietary data is usually e
xploited for commercial advantages
1140 and profit within the organization, and not avail
able to the broad scientific com-
1141 munity. To ground a methodology to model social p
henomena on the availability of
1142 proprietary data that is not in control of the po
licy makers raises few concerns on
1143 the actual implementation in a real scenario.
1144 7 Conclusions
1145 In this paper we proposed 34 di↵erent characteris
tics of individual socio-economic
1146 behavior quantifiable through the dataset of anon
ymized insurance customers, and
1147 then evaluated them on the example of Swiss munic
ipalities. We showed that those
1148 quantities could be used for estimating economic
performance of the regions in
the country, as proposed geographical regression
1149 models technique
demonstrated

1150 to
perform well on the validation samples for predic
ting major ocial
statistical
1151 quantities
for di↵erent categories such as Population, Tran
sportation, Work,
Space
1152 and Territory, Housing and Economy on
the level of Swiss municipalities. Moreover,
1153

1154 Massaroetal.Page 17 of29

1155 the same approach
is applicable in cases when ocial statistics is
not available or is
1156 inconsistent, for example when considering geogra
phical units of a finer spatial scale

1197 Donadioetal.Page 19 of30
1198 heterogeneity of social processes at a micro-leve
l, e.g., neighborhoods in cities.
1199 Temporal evolution. In our analysis we currently
focus on a static snapshot cov-
1200 ering a year of statistics. However, socioeconomi
c conditions vary over time and in
1201 which extent and how fast this change is reflecte
d into the insurance data records
1202 is something not explored yet.
1203 Data availability and privacy. The current approa
ch is based on the assumption
1204 that customers data is available to the researche
rs to tackle relevant challenges that
1205 have a broad social impact. This raises two main
issues related to privacy and the

1206 compliance to the current legislation especially
in the European framework, and
1207 the sharing policy. Proprietary data is usually e
xploited for commercial advantages
1208 and profit within the organization, and not avail
able to the broad scientific com-
1209 munity. To ground a methodology to model social p
henomena on the availability of
1210 proprietary data that is not in control of the po
licy makers raises few concerns on
1211 the actual implementation in a real scenario.
1212 7 Conclusions
1213 In this paper we proposed 34 di↵erent characteris
tics of individual socio-economic
1214 behavior quantifiable through the dataset of anon
ymized insurance customers, and
1215 then evaluated them on the example of Swiss munic
ipalities. We showed that those
1216 quantities could be used for estimating economic
performance of the regions in
the country, as proposed geographical regression
1217 models technique
demonstrated to
perform well on the validation samples for predic
1218 ting major ocial
statistical quan-

1219 tities
for di↵erent categories such as Population, Tran
sportation, Work,
Space and
1220 Territory, Housing and Economy at
the level of Swiss municipalities. This approach
1221 is applicable in cases when ocial statistics are
not available or they are inconsis-
tent, and the experimental pipeline demonstrated
1222 its ability to reach comparable
performance to a scenario with complete knowledg
1223 e. Advantages
and disadvantages

1224 of local and global spatial regression
models have been discussed extensively, high-

geographical regression models. In
the next steps, we aim at applying the same appro
1159 ach also allows evaluating tem-

poral variation of socio-economic condition perfo
1160 rmance of the Swiss cities, which

1161 is especially useful to study process
of urbanization and gentrification. Finally, the
proposed model can be further employed for estima
1162 ting more specific characteristics
of local quality of life of cities and neighbourh
1163 oods.
1164 Author details
1165 1
1166 HERUS Lab,
1167 ´
1168 Ecole polytechnique f´ed´erale de Lausanne, ENAC, IIE GR C1 455 (Bˆatiment GR) - Station 2, 1015 -
1169 Lausanne, CH.
1170 2
1171 University of Turin,, Corso Svizzera 185, 10149 Turin, Italy.
1172 3
1173 ISI Foundation, Via Chisola 5, 10126
1174 Turin, Italy.
1175 References
1176 1.MacFeely, S.: The big (data) bang: Opportunities and challenges for compiling sdg indicators. Global Policy10,
1177 121–133 (2019)
1178 2.IEAG, U.: A world that counts–mobilising the data revolution for sustainable development. New York: United
1179 Nations (2014)
1180 3.Struijs, P., Braaksma, B., Daas, P.J.: Ocial statistics and big data. Big Data & Society1(1),
1181 2053951714538417 (2014)
1182 4.Pappalardo, L., Pedreschi, D., Smoreda, Z., Giannotti, F.: Using big data to study the link between human
1183 mobility and socio-economic development. In: 2015 IEEE International Conference on Big Data (Big Data), pp.
1184 871–878 (2015). IEEE
1185 5.Antenucci, D., Cafarella, M., Levenstein, M., R´e, C., Shapiro, M.D.: Using social media to measure labor
1186 market flows. Technical report, National Bureau of Economic Research (2014)
1187 6.Watmough, G.R., Marcinko, C.L., Sullivan, C., Tschirhart, K., Mutuo, P.K., Palm, C.A., Svenning, J.-C.:
1188 Socioecologically informed use of remote sensing data to predict rural household poverty. Proceedings of the
1189 National Academy of Sciences116(4), 1213–1218 (2019)

1227 the modeling of temporal variations, which
is especially useful to study processes
1228 of urbanization and gentrification.
We also aim at developing models for estimating
attributes at finer geographical resolutions such
1229 as districts or neighborhoods.
1230 Author details
1231 1
1232 HERUS Lab,
1233 ´
1234 Ecole polytechnique f´ed´erale de Lausanne, ENAC, IIE GR C1 455 (Bˆatiment GR) - Station 2, 1015 -
1235 Lausanne, CH.
1236 2
1237 University of Turin,, Corso Svizzera 185, 10149 Turin, Italy.
1238 3
1239 ISI Foundation, Via Chisola 5, 10126
1240 Turin, Italy.
1241 References
1242 1.MacFeely, S.: The big (data) bang: Opportunities and challenges for compiling sdg indicators. Global Policy10,
1243 121–133 (2019)
1244 2.IEAG, U.: A world that counts–mobilising the data revolution for sustainable development. New York: United
1245 Nations (2014)
1246 3.Struijs, P., Braaksma, B., Daas, P.J.: Ocial statistics and big data. Big Data & Society1(1),
1247 2053951714538417 (2014)
1248 4.Pappalardo, L., Pedreschi, D., Smoreda, Z., Giannotti, F.: Using big data to study the link between human
1249 mobility and socio-economic development. In: 2015 IEEE International Conference on Big Data (Big Data), pp.
1250 871–878 (2015). IEEE
1251 5.Antenucci, D., Cafarella, M., Levenstein, M., R´e, C., Shapiro, M.D.: Using social media to measure labor
1252 market flows. Technical report, National Bureau of Economic Research (2014)
1253
1254 Donadioetal.Page 20 of30
1255 6.Watmough, G.R., Marcinko, C.L., Sullivan, C., Tschirhart, K., Mutuo, P.K., Palm, C.A., Svenning, J.-C.:
1256 Socioecologically informed use of remote sensing data to predict rural household poverty. Proceedings of the
1257 National Academy of Sciences116(4), 1213–1218 (2019)

- 733 Removals + 810 Additions

1192 8.Giannone, D., Reichlin, L., Small, D.: Nowcasti
ng: The real-time informational content of macroe
conomic data.
1193 Journal of Monetary Economics55(4), 665–676 (200
8)
1194 9.Ginsberg, J., Mohebbi, M.H., Patel, R.S., Bramm
er, L., Smolinski, M.S., Brilliant, L.: Detecting
influenza
1195 epidemics using search engine query data. Nature4
57(7232), 1012 (2009)
1196 10.Zhou, K., Yang, S.: Understanding household en
ergy consumption behavior: The contribution of en
ergy big
1197 data analytics. Renewable and Sustainable Energy
Reviews56,810–819(2016)
1198 11.De Montjoye, Y.-A., Radaelli, L., Singh, V.K.,
etal.:Uniqueintheshoppingmall:Onthereidentifiabil
ityof
1199 credit card metadata. Science347(6221), 536–539
(2015)
1200 12.Ritter, J., Mayer, A.: Regulating data as prop
erty: a new construct for moving forward. Duke L.

& Tech. Rev.
1201 16,220(2017)
1202 13.Zhang, Y., Chen, X., Li, J., Wong, D.S., Li,
H., You, I.: Ensuring attribute privacy protecti
on and fast
1203 decryption for outsourced data security in mobile
cloud computing. Information Sciences379,42–61(20
17)
1204 14.Sivarajah, U., Kamal, M.M., Irani, Z., Weerakk
ody, V.: Critical analysis of big data challenges
and analytical
1205 methods. Journal of Business Research70,263–286(2
017)
1206 15.Florescu, D., Karlberg, M., Reis, F., Del Cast
illo, P.R., Skaliotis, M., Wirthmann, A.: Will 'b
ig data'transform
1207 ocial statistics. In: European Conference on the
QualityofOcial Statistics. Vienna, Austria, pp.
2–5 (2014)
1208 16.Schnore, L.F.: The socio-economic status of ci
ties and suburbs. American Sociological Review, 7
6–85 (1963)
1209 17.Dennison, G., Dodd, V., Whelan, B.: A socio-ec
onomic based survey of household waste characteri
stics in the
1210 city of dublin, ireland. i. waste composition. Re
sources, Conservation and Recycling17(3), 227–244
(1996)
1211 18.Cameron, S.: The economics of crime deterrenc
e: A survey of theory and evidence. Kyklos41(2),
301–323
1212 (1988)
1213 19.Galea, S., Freudenberg, N., Vlahov, D.: Cities
and population health. Social science & medicine6
0(5),
1214 1017–1033 (2005)
1215 20.Clark, W.A.: Residential segregation in americ
an cities: A review and interpretation. Populatio

1260 8.Giannone, D., Reichlin, L., Small, D.: Nowcasti
ng: The real-time informational content of macroe
conomic data.
1261 Journal of Monetary Economics55(4), 665–676 (200
8)
1262 9.Ginsberg, J., Mohebbi, M.H., Patel, R.S., Bramm
er, L., Smolinski, M.S., Brilliant, L.: Detecting
influenza
1263 epidemics using search engine query data. Nature4
57(7232), 1012 (2009)
1264 10.Zhou, K., Yang, S.: Understanding household en
ergy consumption behavior: The contribution of en
ergy big
1265 data analytics. Renewable and Sustainable Energy
Reviews56,810–819(2016)
1266 11.De Montjoye, Y.-A., Radaelli, L., Singh, V.K.,
etal.:Uniqueintheshoppingmall:Onthereidentifiabil
ityof
1267 credit card metadata. Science347(6221), 536–539
(2015)
1268 12.Ritter, J., Mayer, A.: Regulating data as prop
erty: a new construct for moving forward. Duke L.

& Tech. Rev.
1269 16,220(2017)
1270 13.Zhang, Y., Chen, X., Li, J., Wong, D.S., Li,
H., You, I.: Ensuring attribute privacy protecti
on and fast
1271 decryption for outsourced data security in mobile
cloud computing. Information Sciences379,42–61(20
17)
1272 14.Sivarajah, U., Kamal, M.M., Irani, Z., Weerakk
ody, V.: Critical analysis of big data challenges
and analytical
1273 methods. Journal of Business Research70,263–286(2
017)
1274 15.Florescu, D., Karlberg, M., Reis, F., Del Cast
illo, P.R., Skaliotis, M., Wirthmann, A.: Will 'b
ig data'transform
1275 ocial statistics. In: European Conference on the
QualityofOcial Statistics. Vienna, Austria, pp.
2–5 (2014)
1276 16.Schnore, L.F.: The socio-economic status of ci
ties and suburbs. American Sociological Review, 7
6–85 (1963)
1277 17.Dennison, G., Dodd, V., Whelan, B.: A socio-ec
onomic based survey of household waste characteri
stics in the
1278 city of dublin, ireland. i. waste composition. Re
sources, Conservation and Recycling17(3), 227–244
(1996)
1279 18.Cameron, S.: The economics of crime deterrenc
e: A survey of theory and evidence. Kyklos41(2),
301–323
1280 (1988)
1281 19.Galea, S., Freudenberg, N., Vlahov, D.: Cities
and population health. Social science & medicine6
0(5),
1282 1017–1033 (2005)
1283 20.Clark, W.A.: Residential segregation in americ
an cities: A review and interpretation. Populatio

- 733 Removals + 810 Additions

453(7196), 779 (2008)
22.Podobnik, V., Ackermann, D., Grubisic, T., Lovrek, I.: Web 2.0 as a foundation for social media marketing:
global perspectives and the local case of croatia. In: Cases on Web 2.0 in Developing Countries: Studies on
Implementation, Application, and Use, pp. 342–379. IGI Global, ??? (2013)
23.Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C.: Quantifying the benefits of vehicle
pooling with shareability networks. Proceedings of the National Academy of Sciences111(37), 13290–13294
(2014)
24.Sobolevsky, S., Massaro, E., Bojic, I., Arias, J.M., Ratti, C.: Predicting regional economic indices using big data
of individual bank card transactions. In: 2017 IEEE International Conference on Big Data (Big Data), pp.
1313–1318 (2017). IEEE

Massaroetal.Page 18 of29

25.Fatehkia, M., Kashyap, R., Weber, I.: Using facebook ad data to track the global digital gender gap. World
Development107,189–209(2018)
26.Zagheni, E., Weber, I., Gummadi, K., et al.: Leveraging facebook's advertising platform to monitor stocks of
migrants (2017)
27.Dubois, A., Zagheni, E., Garimella, K., Weber, I.: Studying migrant assimilation through facebook interests. In:
International Conference on Social Informatics, pp. 51–60 (2018). Springer
28.Llorente, A., Garcia-Herranz, M., Cebrian, M., Moro, E.: Social media fingerprints of unemployment. PloS one
10(5), 0128692 (2015)
29.Cesare, N., Grant, C., Hawkins, J.B., Brownstein, J.S., Nsoesie, E.O.: Demographics in social media data for
public health research: does it matter? arXiv preprint arXiv:1710.11048 (2017)
30.Weber, I., Castillo, C.: The demographics of web search. In: Proceedings of the 33rd International ACM SIGIR
Conference on Research and Development in Information Retrieval, pp. 523–530 (2010). ACM
31.Weber, C., Johnson, M., Arceneaux, K.: Genetics, personality, and group identity. Social science quarterly
92(5), 1314–1337 (2011)
32.Augustin, M., Reich, K., Glaeske, G., Schaefer, I., Radtke, M.: Co-morbidity and age-related prevalence of
psoriasis: analysis of health insurance data in g

453(7196), 779 (2008)
22.Podobnik, V., Ackermann, D., Grubisic, T., Lovrek, I.: Web 2.0 as a foundation for social media marketing:
global perspectives and the local case of croatia. In: Cases on Web 2.0 in Developing Countries: Studies on
Implementation, Application, and Use, pp. 342–379. IGI Global, ??? (2013)
23.Santi, P., Resta, G., Szell, M., Sobolevsky, S., Strogatz, S.H., Ratti, C.: Quantifying the benefits of vehicle
pooling with shareability networks. Proceedings of the National Academy of Sciences111(37), 13290–13294
(2014)
24.Sobolevsky, S., Massaro, E., Bojic, I., Arias, J.M., Ratti, C.: Predicting regional economic indices using big data
of individual bank card transactions. In: 2017 IEEE International Conference on Big Data (Big Data), pp.
1313–1318 (2017). IEEE
25.Fatehkia, M., Kashyap, R., Weber, I.: Using facebook ad data to track the global digital gender gap. World
Development107,189–209(2018)
26.Zagheni, E., Weber, I., Gummadi, K., et al.: Leveraging facebook's advertising platform to monitor stocks of
migrants (2017)
27.Dubois, A., Zagheni, E., Garimella, K., Weber, I.: Studying migrant assimilation through facebook interests. In:
International Conference on Social Informatics, pp. 51–60 (2018). Springer
28.Llorente, A., Garcia-Herranz, M., Cebrian, M., Moro, E.: Social media fingerprints of unemployment. PloS one
10(5), 0128692 (2015)
29.Cesare, N., Grant, C., Hawkins, J.B., Brownstein, J.S., Nsoesie, E.O.: Demographics in social media data for
public health research: does it matter? arXiv preprint arXiv:1710.11048 (2017)
30.Weber, I., Castillo, C.: The demographics of web search. In: Proceedings of the 33rd International ACM SIGIR
Conference on Research and Development in Information Retrieval, pp. 523–530 (2010). ACM
31.Weber, C., Johnson, M., Arceneaux, K.: Genetics, personality, and group identity. Social science quarterly
92(5), 1314–1337 (2011)
32.Augustin, M., Reich, K., Glaeske, G., Schaefer, I., Radtke, M.: Co-morbidity and age-related prevalence of
psoriasis: analysis of health insurance data in g

Analysis of statutory health insurance data. Wound Repair and Regeneration24(2), 434–442 (2016)
34.Rawte, V., Anuradha, G.: Fraud detection in health insurance using data mining techniques. In: 2015
International Conference on Communication, Information & Computing Technology (ICCICT), pp. 1–5 (2015).
IEEE
35.Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y.: Auto insurance fraud detection using unsupervised spectral
ranking for anomaly. The Journal of Finance and Data Science2(1), 58–75 (2016)
36.Bell, N., Arrington, A., Adams, S.A.: Census-based socioeconomic indicators for monitoring injury causes in the
usa: a review. Injury Prevention21(4), 278–284 (2015)
37.Berkowitz, S.A., Traore, C.Y., Singer, D.E., Atlas, S.J.: Evaluating area-based socioeconomic status indicators
for monitoring disparities within health care systems: Results from a primary care network. Health services
research50(2), 398–417 (2015)
38.Brechb¨uhl, B., M¨uller, L., Tschirren, M.: Statistiques des villes suisses 2017. Technical report, Union des villes
suisses UVS, Swiss Federal Statistical Oce (FSO) (2017)
39.Glaeser, E.: Cities, productivity, and quality of life. Science333(6042), 592–594 (2011)
40.Kong, L.: Introduction: culture, economy, policy: trends and developments. Geoforum31(4), 385 (2010)
41.Harrell Jr, F.E.: Regression modeling strategies
42.Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society: Series
B(Methodological)58(1), 267–288 (1996)
43.Schleich, M., Olteanu, D., Ciucanu, R.: Learning linear regression models over factorized joins. In: Proceedings
of the 2016 International Conference on Management of Data, pp. 3–18 (2016). ACM
44.Anselin, L.: Spatial Econometrics vol. 310330. Chap. Fourteen
45.Rey, S.J., Anselin, L.: PySAL: A Python Library of Spatial Analytical Methods. The Review of Regional Studies
37(1), 5–27 (2007)

Analysis of statutory health insurance data. Wound Repair and Regeneration24(2), 434–442 (2016)
34.Rawte, V., Anuradha, G.: Fraud detection in health insurance using data mining techniques. In: 2015
International Conference on Communication, Information & Computing Technology (ICCICT), pp. 1–5 (2015).
IEEE
35.Nian, K., Zhang, H., Tayal, A., Coleman, T., Li, Y.: Auto insurance fraud detection using unsupervised spectral
ranking for anomaly. The Journal of Finance and Data Science2(1), 58–75 (2016)
36.Bell, N., Arrington, A., Adams, S.A.: Census-based socioeconomic indicators for monitoring injury causes in the
usa: a review. Injury Prevention21(4), 278–284 (2015)
37.Berkowitz, S.A., Traore, C.Y., Singer, D.E., Atlas, S.J.: Evaluating area-based socioeconomic status indicators
for monitoring disparities within health care systems: Results from a primary care network. Health services
research50(2), 398–417 (2015)
38.Brechb¨uhl, B., M¨uller, L., Tschirren, M.: Statistiques des villes suisses 2017. Technical report, Union des villes
suisses UVS, Swiss Federal Statistical Oce (FSO) (2017)
39.Glaeser, E.: Cities, productivity, and quality of life. Science333(6042), 592–594 (2011)

Donadioetal.Page 21 of30

40.Kong, L.: Introduction: culture, economy, policy: trends and developments. Geoforum31(4), 385 (2010)
41.Khan, G., Qin, X., Noyce, D.A.: Spatial analysis of weather crash patterns. Journal of Transportation
Engineering134(5), 191–202 (2008). doi:10.1061/(ASCE)0733-947X(2008)134:5(191)
42.MORAN, P.A.P.: NOTES ON CONTINUOUS STOCHASTIC PHENOMENA. Biometrika37(1-2), 17–23
(1950). doi:10.1093/biomet/37.1-2.17. https://academic.oup.com/biomet/article-pdf/37/1-2/17/487420/37-1-2-17.pdf
43.Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM Journal on Scientific
and Statistical Computing7(4), 1307–1330 (1986). doi:10.1137/0907087.https://doi.org/10.1137/0907087
44.Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Statist.32 (2), 407–499 (2004).

improved akaike information criterion. Journal of the Royal Statistical Society: Series B (Statistical
Methodology)60(2), 271–293 (1998)
47.Hoaglin, D.C., Welsch, R.E.: The hat matrix in regression and anova. The American Statistician32(1), 17–22
(1978)
48.Zhang, Y., Zhang, Y.: Exploring the relationship between ridesharing and public transit use in the united states.
International journal of environmental research and public health15(8), 1763 (2018)
49.Horowitz, I.: City Politics and Planning. Routledge, ??? (2017)
50.Van Fan, Y., Perry, S., Klemeˇs, J.J., Lee, C.T.: A review on air emissions assessment: Transportation. Journal
of cleaner production194,673–684(2018)
51.Saboori, B., Sapri, M., bin Baba, M.: Economic growth, energy consumption and co2 emissions in oecd
(organization for economic co-operation and development)'s transport sector: A fully modified bi-directional

el. Ann. Statist.6(2), 461–464 (1978).
doi:10.1214/aos/1176344136
46.Sauerbrei, W., Buchholz, A., Boulesteix, A.–L., Binder, H.: On stability issues in deriving multivariable
regression models. Biometrical Journal57(4), 531–555 (2015). doi:10.1002/bimj.201300222.
https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201300222
47.Heinze, G., Wallisch, C., Dunkler, D.: Variable selection – a review and recommendations for the practicing
statistician. Biometrical Journal60(3), 431–449 (2018). doi:10.1002/bimj.201700067.
https://onlinelibrary.wiley.com/doi/pdf/10.1002/bimj.201700067
48.Chernick, M.R.: Bootstrap Methods: A Guide for Practitioners and Researchers. Wiley Series in Probability and
Statistics. Wiley, ??? (2011). https://books.google.it/books?id=UxDKh5Spwp8C

49.Nogueira, S., Sechidis, K., Brown, G.: On the stability of feature selection algorithms. Journal of Machine
Learning Research18(174), 1–54 (2018)
50.Anselin, L.: Spatial Econometrics vol. 310330. Chap. Fourteen
51.Hurvich, C.M., Simono↵, J.S., Tsai, C.–L.: Smoothing parameter selection in nonparametric regression using an
improved akaike information criterion. Journal of the Royal Statistical Society: Series B (Statistical
Methodology)60(2), 271–293 (1998)
52.Hoaglin, D.C., Welsch, R.E.: The hat matrix in regression and anova. The American Statistician32(1), 17–22
(1978)
53.Zhang, Y., Zhang, Y.: Exploring the relationship between ridesharing and public transit use in the united states.
International journal of environmental research and public health15(8), 1763 (2018)
54.Horowitz, I.: City Politics and Planning. Routledge, ??? (2017)
55.Van Fan, Y., Perry, S., Klemeˇs, J.J., Lee, C.T.: A review on air emissions assessment: Transportation. Journal
of cleaner production194,673–684(2018)
56.Saboori, B., Sapri, M., bin Baba, M.: Economic growth, energy consumption and co2 emissions in oecd
(organization for economic co-operation and development)'s transport sector: A fully modified bi-directional

733 Removals 810 Additions

1285 Author's contributions
1286 E.M. collected the data, analysed the data, developed the models, analysed the results, supervised the research,
1287 wrote the paper. L.D. analysed the data, developed the models. R.S. analysed the results, wrote the paper,
1288 supervised the research. C.R.B. wrote the paper, supervised the research. All authors read and approved the final
1289 manuscript.
1290 Acknowledgements
1291 The authors would like to thank the insurance company La Mobiliere for supporting this research and for providing
1292 the dataset used for the analysis.

1367 Author's contributions
1368 E.M. collected the data, analysed the data, developed the models, analysed the results, supervised the research,
1369 wrote the paper. L.D. analysed the data, developed the models. R.S. analysed the results, wrote the paper,
supervised the research. All authors read and app
1370 roved the
final manuscript.
1371 Acknowledgements
1372 The authors would like to thank the insurance company La Mobiliere for supporting this research and for providing
1373 the anonymized dataset used for the analysis.