Bachelor-Thesis

Fine-Grained Population Estimation

Simon Weidner

16.3.2015

Albert-Ludwigs-Universität Freiburg im Breisgau
Faculty of Engineering
Department of Computer Science
Chair of Algorithms and Data Structures

Bearbeitungszeitraum
16. 12. 2014 – 16. 03. 2015

Gutachter
Dr. Sabine Storandt

Betreuer
Dr. Sabine Storandt

Erklärung

Hiermit erkläre ich, dass ich diese Abschlussarbeit selbständig verfasst habe, keine anderen als die angegebenen Quellen/Hilfsmittel verwendet habe und alle Stellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen wurden, als solche kenntlich gemacht habe. Darüber hinaus erkläre ich, dass diese Abschlussarbeit nicht, auch nicht auszugsweise, bereits für eine andere Prüfung angefertigt wurde.

Datum	Unterschrift

Contents

Abstract

This thesis introduces a new method to estimate population numbers for areas of various sizes. The main idea is to estimate population numbers of buildings by exploiting their environment. For this we use features as the count of nearby schools, supermarkets and many more. We obtain such data from OpenStreetMap, which is a volunteered project providing Google-Maps-like data. State of the art approaches mainly rely on costly three dimensional images and/or census data of the local areas. We use machine learning to predict population values. Therefore, our method can be used without local census data. For evaluation purposes we compared our results to the official census data of various cities, districts and villages. Our results indicate that our method basically works, but requires further research especially to distinguish between different types of houses more precisely.

Zusammenfassung

In dieser Bachelorarbeit stellen wir eine neue Methode vor, die Bevölkerungszahlen frei wählbarer Gebiete schätzt. Die Grundidee ist dabei folgende: Die nähere Umgebung eines Gebäudes sagt viel über die Bevölkerungsdichte eines Gebietes aus. Als Indikatoren verwenden wir beispielsweise Einrichtungen, wie Schulen, aber auch Supermärkte. Die nötigen Informationen dafür beziehen wir von dem freien Projekt OpenStreetMap, welches an Google-Maps erinnert. Andere Schätzmethoden benötigen oft entweder hoch aufwändige dreidimensionale Bilddaten und/oder lokale Zensusdaten. Da wir maschinelles Lernen verwenden, um die Bevölkerungszahlen zu schätzen, sind wir nicht auf lokale Zensusdaten angewiesen. Um unsere Herangehensweise zu überprüfen, haben wir unsere Schätzungen mit offiziellen Bevölkerungszahlen von Städten, Stadtteilen und Ortschaften verglichen. Die Ergebnisse zeigen, dass unsere Methode generell funktioniert, aber noch weitere Forschung von Nöten ist, vor allem bei der Unterscheidung verschiedener Häusertypen.

1. Introduction

1.1. Motivation

Population counts are useful for many purposes. May it be for the planning of commercial expansion or risk management of natural catastrophes.

Censuses have been performed since ancient times. It is and always was connected to a large amount of time and effort. Today registration offices, do most of the work enumerating all denizens. To protect privacy, only large chunks of population data get published, e.g. the number of inhabitants of whole districts, villages or towns.

This thesis provides an approach to estimate the inhabitants of precisely selected areas. As a basis, we used data from OpenStreetMap (OSM) [1], see figure 1. OSM is more than a street graph, it is a free editable map of the world. We process the OSM, trying to find indicators for the population density. Then we try to point out which buildings are inhabitated.

Finally real population data from OpenGeoDB [2] is used. Such population data from whole cities will be distributed among its buildings. As the population data is incomplete, we use machine learning to compensate.

1.2. Course of Action

To begin with, we extract relevant information from the OSM. Relevant is in our case anything, which indicates the quantity or absence of population, e.g. schools, playgrounds, leisure facilities on the one side and industrial areas or office buildings on the other side.

Before estimating the population, we try to improve the given data. First, we classify types of areas, as residential, industrial or other areas. The OSM is largely covered with so called 'landuse' tags, which indicate those types. Some buildings have no connected area type. Those will be assigned with a logistic regression classifier.

In the second step, buildings will be divided in two groups: Residential and not residential. We do this by using building information from OSM as training data. Again, logistic regression is used to classify.

In the third step, we divide residential buildings in single family and appartment buildings. Buildings which are not classified by OSM will be specified with logistic regression.

In the last step the population numbers are fixed. To get a basis for the estimation, we use population numbers from OpenGeoDB. OpenGeoDB provides

[1] http://www.openstreetmap.org/
[2] http://opengeodb.org/

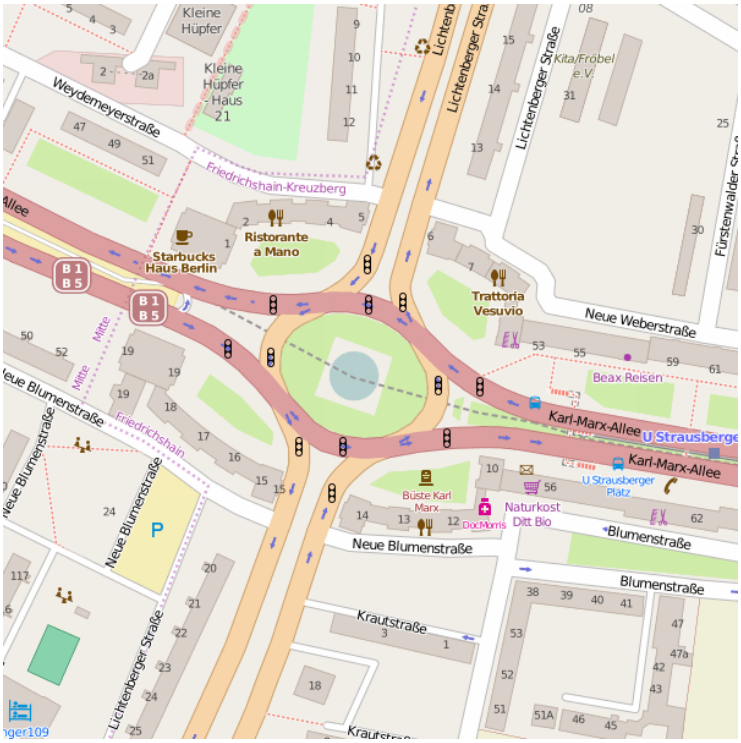

Figure 1: A rendered picture of the OpenStreetMap.

us with population numbers for cities and municipalities. We distribute such population numbers among all residential buildings within the matching borders. Linear regression is used to predict the number of inhabitants for buildings with no provided data.

As part of this thesis, we provide the program FineGrainedPopulationEstimationServer. The program is the population estimator, which will be presented in this thesis. Additionally it has an easy to use web server front end, see figure 2.

With the web server one can select a polygon on a map, which will be evaluated by the program. The program consists of about 3900 lines of code in C++, Python, JavaScript and html. Parsing and interface structures are mostly tested with gtest.

Figure 2: The front end of the estimator.

2. Related Work

There are several pieces of work which aim to estimate the population of particular areas.

2.1. Fine-resolution population mapping using OpenStreetMap points-of-interest

The basic idea in this paper is to seek a correlation between OSM Map Features, and population density. Therefore they [**Bakillahab et al.**, **2014**] use a grid which has a varying size, so that there is a maximum number of map features within each cell. Then they map population data from census blocks to the grid cells. To accomplish a building level resolution, they distribute the data from grid cells among the buildings, with respect to the buildings' proportions. They performed a single, but sophisticated case study about Hamburg with 106 census blocks.

The general difference is that they have very high-resolution population data and they do not share our data improvement steps. Therefore we create with prediction tools new data, where they are solely relying on the given map features. Their approach should be more accurate when evaluating multiple buildings. Our approach takes advantage when deciding if a specific building is populated or not.

Another difference is that we work with sparsely inhabited areas as well as on cities, while they restricted their selves on a single city. They focus on the city Hamburg and we on the whole country of Germany.

2.2. Gridded Population of the World, v3 - Poplation Estimation Service

Gridded Population of the World [**Balk et al.**, **2005**] is a project which creates gridded census data of the whole world [1]. They have a web service quite similar to our front end. One can evaluate an arbitrary polygon. To evaluate they multiply covered area with average population density values. It is rather coarse-grained than fine grained. Each grid cell has a height and length of 2.5 arc minutes, which is about 4 km depending where its measured. Its benefits are that the population values are accurate for the whole earth, though they are always a bit out-dated.

[1] http://sedac.ciesin.columbia.edu/data/collection/gpw-v3/

2.3. ForestMaps: A Computational Model and Visualization for Forest Utilization

Population estimation is only one part of this paper, as they[**Bast et al.**, **2014**] need population data near forests. They use a simple but effective strategy. They accumulate the track length of all streets and weight the sum with the factor one half. The intuition behind this is, as most housings are connected to a street, urban areas have a high street density. As data base they use OSM.

The main advantages of this approach are that it is easy to implement and solely relies on street data. The street coverage is better than the building coverage in OSM. The disadvantage is clearly that it only relies on streets and estimates positive population values, even if there are no buildings nearby.

For evaluation we reimplemented this approach and compared it to our method. In the following we call this method Voronoi-Street-Diagram.

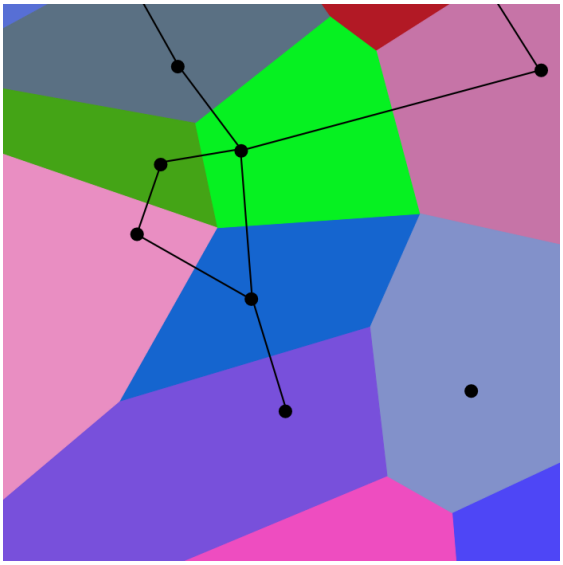

Figure 3: An image of a street graph in black, in colored voronoi cells. The black nodes are the streets vertices.

A Voronoi diagram[1] is a partitioning of a plane into regions based on distance to points in a specific subset of the plane, see figure 3. Their approach is equivalent to computing a Voronoi diagram for all vertices of the streets and summing up the street lengths inside the respective Voronoi cells.

[1]http://en.wikipedia.org/wiki/Voronoi_diagram

2.4. Building population mapping with aerial imagery and GIS data

In this paper, they[**Ural et al.**, **2011**] used data from geographic information systems and aerial images. They used city zoning maps to separate residential areas. Within residential zones they further separated buildings into utility buildings and two kinds of residential buildings. Utility buildings are uninhabited buildings such as garages and barns, but also supermarkets and schools. Residential buildings are divided into apartments and houses. Finally they distributed population values from census blocks with weighted area metric and weighted volumetric models.

$$Population(i) = \frac{weight(i) \cdot S(i)}{\sum_{b \in buildings} weight(b) \cdot S(b)} \cdot totalPopulation$$

Where $S(i)$ is the area or volume of building i according to the chosen model. Both areas and volumes of buildings are extracted from the aerial images.

This paper has many similarities in general, but a completely different realization. They rely on a large set of rules to classify their buildings, we use machine learning. Their approach is very precise, but requires very expensive data as high resolution aerial images, that is why it is not suitable for whole countries.

3. Used Algorithms and Data Structures

3.1. Regression Analysis

Regression Analysis[1] is a statistical process, which estimates a function with the help of observations. The function shall describe the relation between the output variable and the input variables. This process can be used to calculate the conditional expectation of the output, based on the inputs. In the thesis we used two different kinds of regression.

3.1.1. Linear Regression

Linear regression[2] is a statistical method to predict the magnitude of an observation. For linear regression we use the method ordinary least squares. This method finds the one linear function which minimizes the sum of squared vertical distances to all training observations, see figure 4.

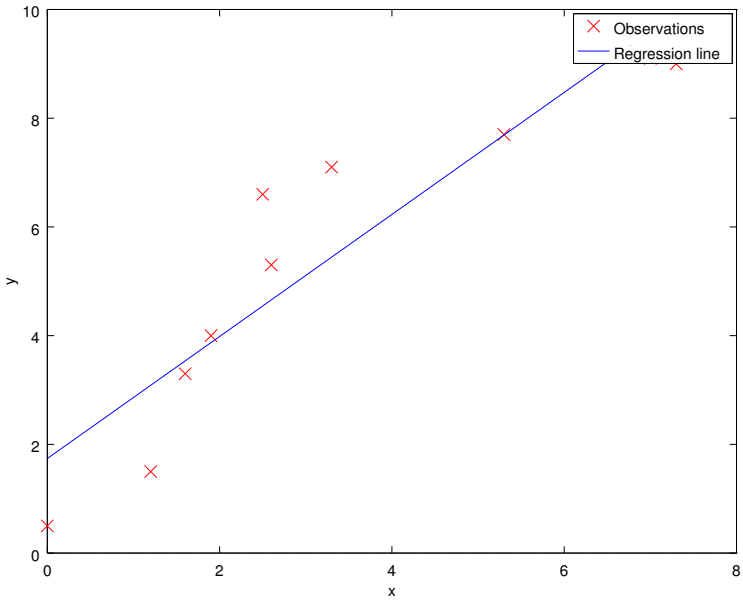

Figure 4: An example of two dimensional linear regression.

Let us have n observations $\{y_1, x_1\}, ..., \{y_n, x_n\}$. $Y = (y_1, ..., y_n)^T$ with $y_i \in \mathbb{R}$ is the observed output and $X = (x_1, ..., x_n)^T$ with $x_i = (x_{i,1}, ...x_{i,m}) \in \mathbb{R}^m$ the

[1] http://en.wikipedia.org/wiki/Regression_analysis
[2] http://en.wikipedia.org/wiki/Linear_regression

observed input. We seek the coefficient vector $\hat{\beta} \in \mathbb{R}^m$, such that $Y = X \cdot \beta + \epsilon$ has the smallest squared error $\|\epsilon\|^2$.

3.1.2. Logistic Regression

Logistic regression[1] is a statistical method that classifies observations into categories. For each category the classifier needs a set of observations. After estimating the influencing value of the input variables, new observations can be classified into one category.

Let us have n observations $X_1, ..., X_n$ with some observations in the category Y_i and some out of Y_i. To model the probability $Pr(Y_i = 1|X)$ we use the logistic function $\delta(t) = \frac{e^t}{e^t+1} = \frac{1}{1+e^{-t}}$ which is monotonically increasing and restricted to values between 0 and 1. As a parameter t we use a linear combination of a input vector X with a coefficient β: $t = \beta_0 + X \cdot \beta$. Thus the probability that an observation is in the category Y_i is $Pr(Y_i = 1|X) = \frac{1}{1+e^{-\beta_0+X\cdot\beta}}$.

We now seek a β which fits our function to the training data $X_1, ..., X_n$. As there is no closed-form solution an approximation is used.

A new observation is classified to $\underset{Y_i}{\text{argmax}}\ Pr(Y_i = 1|X)$.

3.2. R-Tree

An R-Tree[2] is a tree like structure which is used in this thesis for indexing geographic coordinates and polygons. The intuition behind the R-Tree is to keep multiple objects close in the data structure if they are geographically close. The nodes of an R-Tree contain a list of all child nodes and a minimum bounding rectangle. The minimum bounding rectangle is in this case the smallest rectangle which contains all the child nodes.

Queries: We used the R-Tree to find out if something is within a certain rectangle. To do so, one starts at the root node and checks whether the minimum bounding rectangle intersects with the query. Only if it intersects, one has to check each of its child nodes recursively. Every leaf node, we reach this way, might part of the solution and will be checked.

Insertion: To insert a new object there are basically two possibilities. First, the object is fully contained in a subtree, then it can be added to the right node. Second, if the object is not contained within a subtree, then a heuristic is asked, which subtree should be enlarged. If a subtree is chosen which already reached its maximum capacity of child nodes, it is divided and added to the parent node. If

[1] http://en.wikipedia.org/wiki/Logistic_regression
[2] http://en.wikipedia.org/wiki/Rtree

the parent node is exhausted as well, this may propagate up to the root node. In that case a new root node will be created.

4. Fine-Grained Population Estimation

In the following, we will illustrate our approach in detail. At first, we comment on the OSM as source of data and the extraction from it. Followed by the why and hows of refining the data. Finally, we go into detail of distributing available population data and estimating the remainder.

4.1. Data Extraction

Underlying Data

The OSM is a volunteer collaborative project providing freely accessible geographic data. A large benefit is that the complete geographic data is available without legal or technical restrictions. The provided data is of varying quality, but it keeps evolving. The quality varies geographically and according to its creator.

Figure 5: An OSM image with a the bakery *Reis-Bäckerei* without building footprint.

The OSM consists of three structures: Nodes, which basically are points with geographic coordinates. Ways, which are made of multiple nodes. And relations, which are made of nodes, ways and/or relations. Additionally all of those may have tags. Tags are key value pairs of various types. For example a Way could have a tag: *key = building, value = residential house*. Many tags are specified as Map Features [1]in the OSM wiki, but they can be chosen arbitrarily. This can lead

[1]http://wiki.openstreetmap.org/wiki/Map_Features

to various errors. To list a few: Typos, non English or home-brewed tag keys or values. At some places the coverage of building footprints is poor. Affected are primarily small isolated villages, which are often sparsely populated.

An analysis from November 2011 in Germany indicated a completeness of 25%. in the federal states of North Rhine-Westphalia and 15% in Saxony. Although further analyses from 2012 confirm that data completeness in Saxony has risen to 23%, the rate of new data input was slowing in the year 2012. [**Hecht et al.**, **2013**]

The quality of the building footprints varies greatly as well. Many building shapes are simplified, some buildings are combined in one big block. But some others are very sophisticated, adding rails, house numbers, traffic lights, telephone booths and public trash cans. There are shops and restaurants, which have a precise position within a specific building, including on which level it is. Some of the buildings have a tag, for what purpose they are used.

Using OSM data it would be of benefit if the data were complete. The OSM has also tags for areas, for which purpose it is used, but not area-wide. Such areas hold information about localized buildings as well, e.g. in a residential area are mainly residential buildings. If every building had a tag indicating if it is inhabited, we only had to distribute the population among those. But only a small fraction has such a tag. Even better would be more specific information, as the building levels or how many apartments a multi-story house has. As the OSM is incomplete, we try to estimate some of the missing data. During the following steps we estimate, to what kind of area a building belongs and whether it is populated. In the latter case, we also try to predict if its a single or multi-family house.

Some of the OSM building data is broken. There is a small number of misshaped buildings. Reasons may be missing nodes or a wrong order of nodes in the outline of a buildings footprint. We detect such buildings and replace their footprint with a respective bounding rectangle.

Data Extraction

We read the OSM from one single XML file. A dictionary is used to transform the reference from id tags to pointers.

Another dictionary is used to keep all keys and values. In this way, the nodes, ways and relations only need to save references instead of strings.

Some nodes' tags characterize buildings, but lack any reference, other than its position on the map. Without connecting that information, it will be harder guessing if a building is populated. So we insert all nodes with such an information into an R-Tree. Then we query the R-Tree for buildings containing such nodes.

We also want to know whether a building is within a town, village, hamlet or whatsoever. As hamlets have often lack boundary data, we assign buildings to the nearest tagged place.

4.2. Area Classification

OSM users can specify if a region is a forest, meadow, a residential area or something else. We hope such information will give us insight about the population. The basic idea is, houses in residential areas are most probable living houses. Whereas those in industrial areas are not. So we want to know to which kind of area a building belongs.

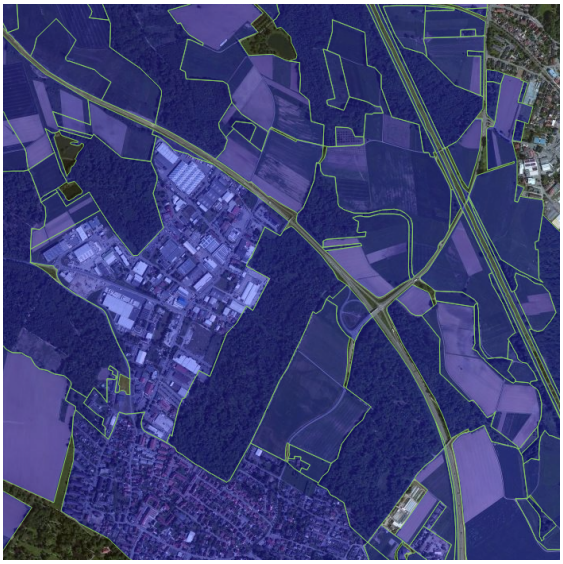

Figure 6: A satellite image, blue areas are specified with the OSM key *landuse*.

Our algorithm has three phases. In phase one, we divide all buildings in two groups, the ones that are in a tagged area and the other consequently. In figure 6, one can see tagged areas. The second phase creates a table with properties of the buildings. The third uses the table to learn from group one and predict the others.

Phase one is straight forward, so we skip to phase two. We simply count how often a specific tag appears in the surrounding of a building, e.g. within x meters. Our aim is to distinguish residential areas from commercial, industrial or other urban areas. We count tags, which appear in such areas, as shops, craft producers, tourist attractions, schools, universities, playgrounds and parks. To count we built an R-Tree for each property. If ways or relations have specific properties, the R-Tree receives only their centroid, for a better running time. Then we query the R-Tree for each building. As query we use a rectangle centered on the building

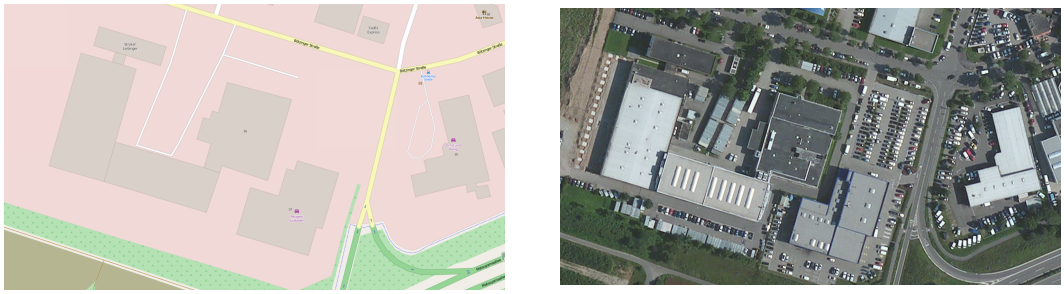

Figure 7: An OSM image and a satellite image of an industrial area. The OSM map does not have tags for every building, though the satellite map indicates that those buildings are industrial.

and filter outputs by distance to the buildings center. We used different distances based on the property. We seek for schools and universities within one to three kilometers, as they have a large area of influence. Leisure facilities, playgrounds and parks, are often located within residential areas. Thus we look out for such places within 50 to several hundred meters. Same for small shops, offices or craft producers, but they might indicate commercial areas. Additionally we counted how many buildings are within the neighborhood, as a measurement of building density.

In the last phase, a logistic regression classifier uses the first group to classify the others.

4.3. Building Classification

At first, we want to distinguish living houses from other buildings. There after, we seek a measurement of how many inhabitants a building has. Most buildings, which are not officially used have not any tags. Only a tiny fraction of buildings have a tag describing its height. In the second step, we separate densely populated apartment buildings from sparsely populated single family houses.

We will describe how we distinguish living houses from other buildings first. In figure 7, you can see an example why only evaluating tags would not work as well. OSM information about buildings will be used to classify a fraction of all buildings. Therefore we identify living houses with tags like: *building = residential* or *building = apartment*. Whereas not residential houses have tags like: *building = industrial*, *building = civic* or *building = hospital*. To supplement the data, we classify buildings as inhabited, if they are within residential areas and have no tags indicating a business of any kind.

As we did for the area classification, we create a table by counting tags near every building. Additionally we create a column for the area, where building resides and its base area. This time we seek information about a single building and not

anymore about the area a building is in. So the range where we seek for tags is smaller than during the area classification. We now seek universities and schools within several hundred meters, instead of kilometers.

To divide the living houses into single and multi family houses, we used a similar approach as before. In this classification part, we want to differentiate between differently dense populated buildings.

We feed our classifier with the type of locality it is in. We differentiate between hamlets, villages, suburbs and towns. The idea is, villages and hamlets have less multi-story dwellings. Another difference is that we incorporate malls and supermarkets separated from other shops. Additionally, we seek for some tags in close and in wide distance ranges. Wide distance properties rather describe the area a building resides. Where small distances are specifying the building. For instance in a village or small city there might be a supermarket next to some single family house, but within a city its more probable that a supermarket will be constructed close to highly populated buildings.

4.4. Population Distribution and Estimation

Lastly, we map population numbers from large areas to the localized living houses. The largest areas we took are whole cities. Therefore population numbers from OpenGeoDB are used. Additionally we take administrative boundary data from OSM, see figure 8. Finally we connect those two parts of data.

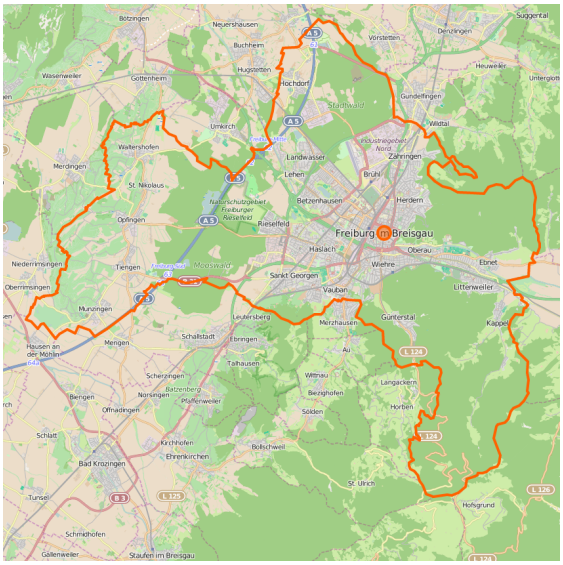

Figure 8: A picture of a boundary taken from OSM.

Now we distribute every population number among the buildings within the matching boundary. If we have different resolution data, for a certain area, we distribute the most high-resolution data first. Then, we remove its population and buildings from all boundaries which contain that one boundary and continue. Distribution will be done with a weighted area metric model from [**Ural et al.**, **2011**].

$$Population(i) = \frac{weight(i) \cdot area(i)}{\sum_{b \in buildings} weight(b) \cdot area(b)} \cdot totalPopulation$$

To determine the weight we use multiple parameters. Uninhabited houses have a weight of zero. The type of building, whether it is a single family house or an apartment building is one parameter. A building gains a penalty, if a business is located within the building. As we had virtually no training data for villages, we tried to predict suitable numbers by incorporating them in our weighting scheme. A penalty is given for buildings within hamlets and villages, because they are not as densely populated as in towns.

After distributing population for known areas, we create a model with linear regression. The idea behind this is, that the number of specific map features give insight about the population numbers. For example if in an area with a size of one square kilometer had five supermarkets and three different schools, it is very probable that a lot of families reside there. On the other side, if there is only one supermarket and a single school, you probably found a sparsely populated village. Using multiple indicators for population density gives us a good estimation whether an area is densely populated. Again we use for our estimation tag density for shops, craft producers, tourist attractions, schools, playgrounds and parks at different distances. To find the right population number for a certain building, we incorporate properties of the building itself as its base area and if it is a single-family building or a multi-family building. We do not use information about the locality, except for hamlets. This is because we want to estimate same values for villages and small towns, if they are built up the same way. Very small settlements as hamlets are an exception, because they are always populated even if there is nothing except farmland nearby. Lastly, we seek farmland in a wide area, as only highly populated urban agglomerations have no farmland within several kilometers.

4.5. Web Server

The Web Server is a small application which answers requests with the readied estimation. A user can specify a custom polygon with simple mouse clicks on one

of two scalable maps of Germany. One map is from OpenStreetMap and the other is a satellite map from Bing[1]. Evaluation is done in real-time.

[1] http://www.bing.com/maps/

5. Evaluation

In the evaluation part we will proceed as in our main part. At first we evaluate the area classification step, followed by the building classification steps. Finally, we will compare the population estimation results to census data and another approach.

5.1. Area Classification Results

As training data, we have 17.3 million buildings which are within tagged areas and only 1.7 million are not. So the OpenStreetMap has a quite good coverage of land use tags. To rate our estimator, we took 50000 buildings out of the training set and predicted on them. The precision was 83%.

5.2. Building Classification Results

To differentiate between inhabited and uninhabited houses, we used 60 OSM tags, see Appendix A. Additionally we claim that all buildings without special tags within residential areas are residential. The OSM of Germany has 18.9 million buildings. Using our classification 14.6 million are inhabited, 1.1 million are uninhabited, the remainder is not classified.

To evaluate these results, we removed 50000 buildings from the training set and predicted on them. The precision was 90%. In addition, we evaluated many different areas and buildings with the following results: Buildings within industrial areas are almost never populated. Hamlets and farms are populated. Multi part buildings, may be partly populated, e.g. a shopping mall, which has multiple parts for different shops. Some parts are not tagged and might be populated, because they are within a residential area. Though, a terraced housing might suffer the complementary problem.

Key	Value	Single Family House
building	apartments	no
building	apartment	no
building	house	yes
building	detached	yes
building	dormitory	no
building	houseboat	yes
building	static_caravan	yes

Table 1: A table with OSM tags indicating whether the building is a single or multi family house.

To differentiate between single family and apartment houses, we used solely seven tags, see 1. Data from the Federal Statistical Office of Germany [1] indicated that 66% of all living houses are single-family houses. The other are two or more family houses and hostels.

Using those seven tags to classify buildings from the OSM, we gain one million classified buildings. 88% of those are single-family houses. After our prediction, we assume that only 6% of all residential buildings are multi-family houses. That result is not very pleasing. This will lead into overestimating the number of areas poor in multi-family houses, while underestimating areas rich in multi-family houses.

5.3. Population Estimation Results

Evaluating the whole process, we compared our results to census data. Additionally, we compared it to the Voronoi-Street-Diagram approach[**Hannah et al.**, **2014**]. This approach will serve as our baseline. We took samples for cities, districts and villages of various sizes.

We did not evaluate certain large cities as whole, because they were used for training, see Appendix B. The census data was taken from UNdata. Our approach works better than the baseline for large and medium-sized cities, see tables 2 and 3. Though there are some outliers, especially Oldenburg and Celle have by far too high numbers, Hanover and Münster too low. One reason might be that both Oldenburg and Celle have an old city core and thus many single family houses. Hanover and Münster on the other side have been largely rebuilt last century and may have more multi family houses than estimated.

For small regions our method is at least on a par with the baseline, see 4.

Our algorithm frequently predicts too high population numbers. For Germany as whole it predicts 80.77 million inhabitants, correct would be 80.62 million. This seems very good, though most places are still overestimated, because the OSM's lack of building data at other places. Main reasons might be that we have training data mostly from highly populated cities and that the second building classification part does not work well.

[1] https://www.destatis.de/

Region	Method	Population	Absolute Error	Relative Error
Essen				
	Estimation	720 280	146 812	26%
	Baseline	129 031	-444 437	-77%
	Census (2012)	573 468		
Hanover				
	Estimation	440 522	-82 092	-16%
	Baseline	145 427	-380 448	-72%
	Census (2012)	525 875		
Münster				
	Estimation	264 433	-27 321	-9%
	Baseline	96 179	-195 575	-67%
	Census (2012)	291 754		
Aachen				
	Estimation	280 298	21 634	8%
	Baseline	79 194	-179 470	-69%
	Census (2012)	258 664		
Kassel				
	Estimation	238 255	41 759	21%
	Baseline	67 352	-129 144	-65%
	Census (2012)	196 496		
Mülheim an der Ruhr				
	Estimation	171 389	4 233	3%
	Baseline	50 694	-116 462	-70%
	Census (2011)	167 156		
Oldenburg				
	Estimation	305 785	146 175	92%
	Baseline	49 011	-110 599	-69%
	Census (2013)	159 610		
Ingolstadt				
	Estimation	174 260	45 124	35%
	Baseline	49 408	-79 728	-62%
	Census (2013)	129 136		
Ulm				
	Estimation	130 069	6 397	-5%
	Baseline	65 160	-58 512	-47%
	Census (2012)	123 672		

Table 2: Estimation results for large German cities.

Region	Method	Population	Absolute Error	Relative Error
Marburg				
	Estimation	87 416	6 269	8%
	Baseline	47 969	-33 178	-41%
	Census (2013)	81 147		
Giessen				
	Estimation	103 079	25 346	33%
	Baseline	39 518	-38 215	-49%
	Census (2013)	77 733		
Celle				
	Estimation	114 762	46 254	68%
	Baseline	54 105	-14 403	-21%
	Census (2013)	68 508		
Buxtehude				
	Estimation	38 527	-1 250	-3%
	Baseline	19 596	-20 181	-51%
	Census (2013)	39 777		
Emmendingen				
	Estimation	34 259	7 972	30%
	Baseline	11 745	-14 542	-55%
	Census (2013)	26 287		
Quickborn				
	Estimation	18 421	-1 621	-8%
	Baseline	10 985	9 057	45%
	Census (2013)	20 042		
Teningen				
	Estimation	14 712	3 167	27%
	Baseline	10 794	-751	-7%
	Census (2013)	11 545		
Staufenberg				
	Estimation	11 120	2 973	36%
	Baseline	9 067	920	11%
	Census (2013)	8 147		
Simonswald				
	Estimation	4 936	1 929	64%
	Baseline	14 922	11 915	397%
	Census (2013)	3 007		

Table 3: Estimation results for medium-sized and small German cities.

Region	Method	Population	Absolute Error	Relative Error
Misburg-Anderten (Hannover)	Estimation	33 853	1 324	4%
	Baseline	11 062	-21 467	-66%
	Census (2014)	32 529		
Puchheim	Estimation	18 605	-1 919	-9%
	Baseline	4 962	-15 562	-76%
	Census (2013)	20 524		
Stühlinger (Freiburg)	Estimation	21 134	2 818	15%
	Baseline	2 556	15 760	-86%
	Census (2014)	18 316		
Donnerschwee (Oldenburg)	Estimation	10 662	3 204	43%
	Baseline	1 665	-5 793	-77%
	Census (2014)	7 458		
Ihringen	Estimation	6 023	133	2%
	Baseline	9 203	3 313	56%
	Census (2013)	5 890		
Mömlingen	Estimation	4 202	-644	-13%
	Baseline	5 257	441	8%
	Census (2013)	4 846		
Zoo (Hannover)	Estimation	9 268	4 639	100%
	Baseline	5 387	758	16%
	Census (2013)	4 629		
Opfingen (Freiburg)	Estimation	2 211	-2 229	-50%
	Baseline	5 655	1 215	27%
	Census (2013)	4 440		
Poltringen	Estimation	2 207	427	24%
	Baseline	2 247	467	26%
	Census (2011)	1 780		

Table 4: Estimation results for villages and urban districts.

6. Future Work

6.1. More Precise Building Classification

As already mentioned we estimate a too low amount of multi-family houses. One approach to fixing this problem would be to change the training data. We could use more data about the buildings outline. Very large quadratic or rectangular shaped residential buildings are very probable to be inhabited by a large number of persons. Single family houses are often detached and are almost never connected with footways from multiple sides.

6.2. Improve the Population Distribution

One flaw of the current implementation is that we lack training data for small areas as villages or city districts. Currently we connect boundaries with the exact same name and same NUTS [1] category. Small differences in the notation already deny a possible connection. We might train a system, which uses more rules to connect locations to boundaries. Words and abbreviated forms with the same meaning should be merged. Different data bases could be used to find more correlations as well.

6.3. Distribute Population Within Areas

In many small villages the OSM lacks buildings, see figure 9. This leads to wrong estimation results.

Figure 9: An OSM image with a village, without any buildings. Though the grey area is a residential area.

[1] http://en.wikipedia.org/wiki/Nomenclature_of_Territorial_Units_for_Statistics

A solution to this problems might be to learn population data for areas first. Then populate buildings within each area. If the number of buildings is oddly small, the population data should not be distributed. However there should be an option, which allows to access to areal population values, for such areas.

Another idea would be using the Voronoi-Street-Diagrams to estimate such areas, as it works quite well for villages.

Danksagung

An dieser Stelle möchte ich mich bei all jenen bedanken, die mich während des Zeitraums der Bachelor Arbeit motiviert und unterstützt haben.

Ein besonderer Dank gilt dabei meiner Betreuerin Dr. Sabine Storandt, welche mich für das Thema der Arbeit begeistern konnte, während des Zeitraums motiviert hat und stets mit Rat beiseite stand.

Weiterhin möchte ich mich bei meiner Familie und meinen Freunden für ihre Unterstützung bedanken.

Zu guter Letzt möchte ich mich bei meinen Korrekturlesern Joanna Probst und Dr. Jan Weidner bedanken.

Vielen Dank euch allen!

Appendices

A. Inhabitation Classification of OSM Data

Key	Value	Inhabited
amenity	prison	yes
building	apartments	yes
building	apartment	yes
building	house	yes
building	detached	yes
building	residential	yes
building	dormitory	yes
building	houseboat	yes
building	static_caravan	yes
landuse	residential	yes
landuse	commercial	no
landuse	industrial	no
amenity	school	no
amenity	college	no
amenity	kindergarten	no
amenity	library	no
amenity	university	no
amenity	fuel	no
amenity	clinic	no
amenity	hospital	no
amenity	cinema	no
amenity	community_centre	no
amenity	theatre	no
amenity	courthouse	no
amenity	crypt	no
amenity	fire_station	no
amenity	firestation	no
amenity	police	no
amenity	waste_disposal	no
amenity	waste_dump	no
building	commercial	no
building	industrial	no
building	retail	no
building	warehouse	no
building	chapel	no
building	church	no
building	hotel	no
building	mosque	no
building	temple	no
building	civic	no
building	hospital	no
building	school	no
building	train_station	no
building	transportation	no
building	university	no
building	public	no
building	barn	no
building	bridge	no
building	bunker	no
building	construction	no
building	garage	no
building	garages	no
building	greenhouse	no
building	hangar	no
building	roof	no
building	sty	no
office	administrative	no
shop	mall	no
shop	shopping_centre	no

B. Regions Used as Training Data for the Population Estimation

Norderney	Aschaffenburg
Nürnberg	Schweinfurt
Flensburg	Düsseldorf
Kiel	Bremen
Lübeck	Frankenthal (Pfalz)
Baden-Baden	Köln
Ludwigshafen am Rhein	Darmstadt
Kaufbeuren	Memmingen
Speyer	Chemnitz
Potsdam	Passau
Mülheim an der Ruhr	Mainz
Landau in der Pfalz	Osnabrück
Frankfurt am Main	Bottrop
Erlangen	Straubing
Rostock	Halle (Saale)
Augsburg	Bayreuth
Mönchengladbach	Pirmasens
Regensburg	Bochum
Delmenhorst	Bielefeld
Wolfsburg	Leipzig
München	Saarbrücken
Cottbus	Kaiserslautern
Wilhelmshaven	Bremerhaven
Leverkusen	Salzgitter
Suhl	Gera
Worms	Mannheim
Remscheid	Jena
Duisburg	Offenbach am Main
Würzburg	Solingen
Brandenburg an der Havel	Kempten (Allgäu)
Pforzheim	Coburg
Wuppertal	Schwabach
Magdeburg	Neustadt an der Weinstraße
Wiesbaden	Oberhausen
Bonn	Erfurt
Gelsenkirchen	Krefeld

Freiburg im Breisgau
Trier
Dresden
Heidelberg
Dortmund
Stuttgart
Gramzow
Gnoien
Willich
Dahlem
Volkmarsen
Bad Arolsen
Admannshagen-Bargeshagen
Eberdingen
Wartenberg
Marienberg
Rosenthal
Adorf
Homberg
Hermsdorf
Homberg
Aurich
Forstinning
Buch
Neuötting
Kösching
Cham
Falkenberg
Möhrendorf
Baiersdorf
Lichtenberg
Homberg
Reinsfeld
Kaulsdorf
Britz
Schöneberg
Blankenburg

Malchow
Dahlem
Finkenthal
Neundorf
Weißensee
Asbach-Sickenberg
Frankfurt (Oder)
Bamberg
Dessau-Roßlau
Braunschweig

References

[**Bakillahab et al., 2014**] Mohamed Bakillahab,Steve Liang, Amin Mobasheria, Jamal Jokar, Arsanjania & Alexander Zipf, *Fine-resolution population mapping using OpenStreetMap points-of-interest*, GIScience Research Group, Institute of Geography, University of Heidelberg, Heidelberg, Germany and Department of Geomatics Engineering, University of Calgary,Calgary, Canada, 18. March 2014.

[**Bast et al., 2014**] Hannah Bast, Jonas Sternisko, and Sabine Storandt, *ForestMaps: A Computational Model and Visualization for Forest Utilization*, Department of Computer Science, University of Freiburg (Germany) {bast,sternis,storandt}@informatik.uni-freiburg.de, 2014 W2GIS.

[**Balk et al., 2005**] Core-group leaders Deborah Balk, Gregory Yetman and many more, *Gridded Population of the World, Version 3 (GPWv3)*, Center for International Earth Science Information Network (CIESIN), Columbia University; Centro Internacional de Agricultura Tropical (CIAT) 2005.

[**Hecht et al., 2013**] Robert Hecht, Carola Kunze and Stefan Hahmann, *Measuring Completeness of Building Footprints in OpenStreetMap over Space and Time*, Leibniz Institute of Ecological Urban and Regional Development Germany, Institute for Cartography, Dresden University of Technology, SPRS Int. J. Geo-Inf. 2013.

[**Ural et al., 2011**] Serkan Ural, Ejaz Hussain, Jie Shan, *Building population mapping with aerial imagery and GIS data*, School of Civil Engineering, Purdue University, West Lafayette, IN 47907, USA, 2011.