Fine-Grained Population Estimation

Hannah Bast
University of Freiburg
79110 Freiburg, Germany
bast@informatik.uni-freiburg.de

Sabine Storandt
University of Freiburg
79110 Freiburg, Germany
storandt@cs.uni-freiburg.de

Simon Weidner
University of Freiburg
79110 Freiburg, Germany

ABSTRACT

We show how to estimate population numbers for arbitrary user-defined regions, down to the level of individual buildings. This is important for various applications like evacuation planning, facility placement, or traffic estimation. However, census data with precise population numbers is typically only available at the level of cities, villages, or districts, if at all.

Previous approaches either rely on available census data for already small areas or on sophisticated input data like high resolution aerial images. Our framework uses only freely available data, in particular, OpenStreetMap data. In the OpenStreetMap project, crowd-sourced data is collected about street networks, buildings, places of interest as well as all kind of regions and natural structures world-wide. We use this data to learn three classifiers that are relevant for the population distribution inside an area: residential vs. industrial vs. commercial landuse, inhabited vs. uninhabited buildings, and single-family vs. multi-family houses. Once learned, we can use these classifiers for population estimation even in areas without any census data at all.

Our experiments show good average accuracy (measured as the deviation from actual census data) for rural areas (25%), metropolitan areas (10%), and cities in countries other than that containing the training data (12%).

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Measurement

Keywords

Population Estimation, Machine Learning, OpenStreetMap Data, Extrapolation

SIGSPATIAL'15, November 03-06, 2015, Bellevue, WA, USA
 ISBN 978-1-4503-3967-4/15/11...$15.00
DOI: http://dx.doi.org/10.1145/2820783.2820828

Figure 1: Our implementation allows for population estimation inside arbitrary user-defined regions. The left image shows a coarse-grained example (on OSM Tiles), the right image illustrates a query demanding fine-grained population information (visualized on Bing Satellite Map Tiles).

1. INTRODUCTION

Fine-grained population estimates are the basis for various applications like evacuation planning [18], studies on infection spreading [10], facility placement, or traffic estimation [13]. It would be ideal for such applications if population counts were available on the level of individual buildings. This data is indeed recorded by registration offices, but due to privacy issues, only aggregated numbers for larger administrative units (like cities, villages, or districts at best) get published.

We present a framework that provides such fine-grained estimates based on OpenStreetMap (OSM)[1] data. OSM data is freely available and contains information about street networks, buildings, places of interest as well as all kind of regions and natural structures world-wide. We use this data to *learn* features that enable us to estimate the number of inhabitants for individual buildings. Via simple aggregation this allows us to estimate population numbers in arbitrary user-defined regions. We provide a live demo for Germany[2]. Figure 1 provides two screenshots, each at a different level of granularity.

Unlike previous work, our approach does not only allow *interpolation* from available census data, but also *extrapolation* to regions where no such data is available. This is

[1] `http://www.openstreetmap.org`
[2] `http://ad.informatik.uni-freiburg.de/publications`

explained in the following sections.

1.1 Related Work

The prevailing line of attack in previous work on population estimation is by interpolation. It is assumed that population counts are available for a certain (sometimes already fairly small) area. In the GPW project[3], these counts are simply uniformly divided over the (4x4km) cells of a grid. More sophisticated approaches use all kinds of additional information to better estimate the population distribution within a given area.

LandScan [6] is a global population database which interpolates census data to population counts for 1x1km grid cells. They use multiple information sources to obtain a high data quality. The final estimations are not openly available but have to be purchased from the LandScan^{TM} company.

In [12], building heights obtained from LiDAR (laser-based remote sensing) data are used for that purpose. In [17], high-resolution satellite images are used together with building detection and classification tools. Both of these works provide population estimates on the level of individual buildings. However, the used data is expensive to gather and generally not openly available.

In [15], the coarse estimates from the GPW project are refined using satellite night-time light imagery. No explicit accuracy numbers are provided, but correlations between light frequency and population density.

Other features used for interpolation include landuse classification [5], street density [19], or so-called *control points* like schools or hospitals that indicate that many people live nearby [20]. These approaches work by dividing a given area (with known population count) into zones with different population densities. Within such a zone, again a standard distribution is assumed. These methods therefore do not lead to population estimates down to the level of individual buildings. Our new approach uses landuse classification as one of its ingredients; see Figure 2.

In [1], footprints of buildings and points of interest (POIs) are extracted from OSM data, like in our approach. Their areal interpolation proceeds in three main steps. First, landuse classification (from an urban atlas) is used to identify residential areas. Second, residential areas are divided into grid cells and population numbers per cell are estimated using selected POIs as control points. Third, the estimate for each grid cell is divided among the contained buildings proportional to their base area. Unlike our approach, they rely on the availability of census data on an already fine-grained level (districts with about 2,000 inhabitants). Also, evaluation is provided only within a large city (Hamburg, Germany). It was observed in [16], that population estimation methods which work well in metropolitan areas often fail in rural areas (for example, due to sparseness of POIs).

In [2], population counts are also based on OSM data. For a given area with a known population, a Voronoi diagram for all points in the street network is computed which partitions the area into cells. Then to every point a population number is assigned that is proportional to the accumulated length of living streets inside its Voronoi cell. The estimates are further adjusted based on the surrounding street density. This approach works without fine-grained census data and for any kind of area, just like our approach. We therefore use this approach as a baseline in our evaluation. As we see in Section 5, we achieve much improved estimates.

[3]Gridded population of the world: `http://sedac.ciesin.columbia.edu/data/collection/gpw-v3`.

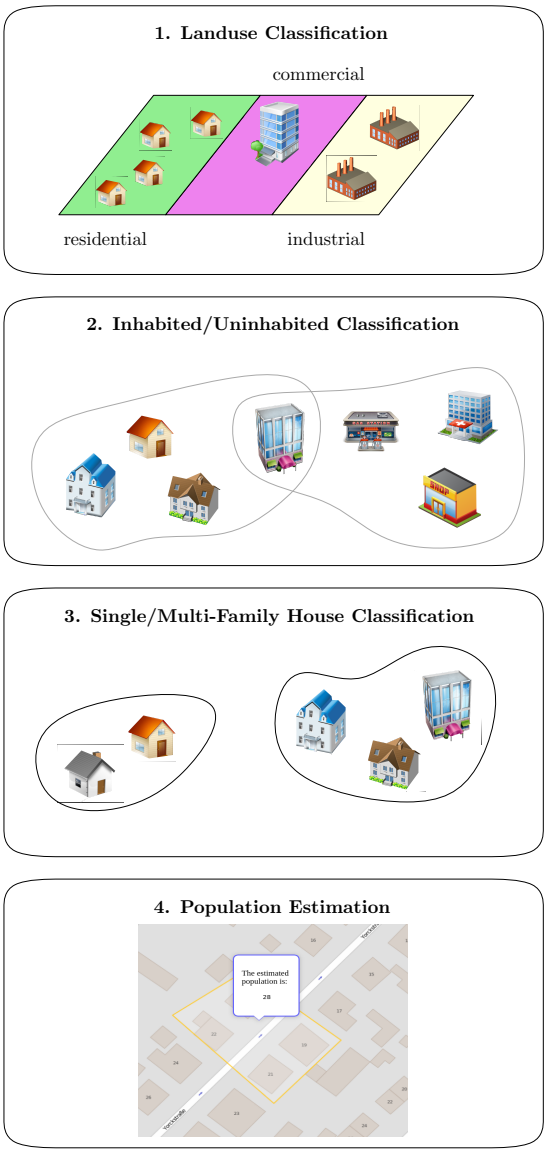

Figure 2: The four main steps in our population estimation framework.

1.2 Contribution and Overview

We consider the following as our main contributions:

- A new approach to population estimation on the level of individual buildings that is based on freely available data (OSM) and does not require fine-grained census data. Unlike previous approaches, our learning-based approach can not only intrapolate from available census data, but it can also extrapolate to regions where no such data is available.
- An evaluation which shows much improved estimates over previous work that is based on freely available data and does not require fine-grained census data. In particular, we achieve good average accuracy (measured as the deviation from actual census data) for rural areas (25%), metropolitan areas (10%), and cities in countries other than that contain-

ing the training data (12%).

• A complete implementation with a working demo (see the link in the introduction).

Our approach works in the following four main steps, which are illustrated in Figure 2.

• *Classification of Landuse.* Like in previous approaches, we use the landuse of an area (residential vs. industrial vs. commercial) as a feature indicative of its population distribution. We develop a classification tool for areas with unknown landuse in OSM.

• *Classification of Inhabited vs. Uninhabited.* In previous work, binary asymmetric mapping [7] was used to subdivide an area into populated and unpopulated cells. We go one step further and also classify buildings as inhabited (e.g. residential houses, apartment buildings) or uninhabited (e.g. schools, shops, restaurants, industrial buildings, office buildings). The landuse, as determined by the previous classifier, is used as one of the features.

• *Classification of Single-Family vs. Multi-Family Houses.* The number of inhabitants per building is clearly correlated with the number of floors. We show how to classify buildings into single-family and multi-family houses when the building height is unknown.

• *Final Population Estimation.* We use all previously gathered information, and population numbers for some selected cities and villages (from OpenGeoDB[4]) to estimate the number of inhabitants per building.

2. DATA EXTRACTION FROM OSM

The goal of the OSM project is to create a free map of the world, by collecting geo-information contributed by volunteers and making it easily accessible for everyone. OSM data is composed of three kinds of elements: *nodes* (as latitude and longitude), *ways* (referencing nodes) and *relations* which are compositions thereof (i.e. referencing sets of nodes, ways or other relations). All three elements can be augmented with tags, which allow to name and classify the data and provide arbitrary additional information. Tags are key-value pairs of various types. For example a way could have the tag: *key = building, value = residential house.* Many tags are specified in the OSM wiki[5], but they can be chosen arbitrarily. We use OSM to extract relevant data for population estimation. Here, relevant refers to anything which indicates the quantity or absence of population. In the following, we provide a detailed overview of the data we extracted from OSM.

- *Building Footprints.* Buildings can be represented in OSM as single nodes or ways/relations if the building footprint is available. At some places the coverage of building footprints is poor. This primarily affects small isolated villages, which are often sparsely populated [8]. Also the quality of the footprints varies. Building shapes might be simplified or several buildings might be combined into one big block. If the building footprint is mis-shaped (that is, lines crossing the interior, e.g., due to missing or wrong nodes), we replace their footprints with the respective bounding rectangle.

- *Building Tags.* Ideally, we would like to have a tag for every building providing the number of its occupants. In the current OSM data, less than 1% of the buildings in Germany exhibit such a tag. For the others, we have to estimate this number. We therefore extract, if available, the building type (e.g., apartment, house, church), the amenity (e.g., kindergarten, theater, police, waste disposal) and look for shop tags (e.g., mall or shopping center). We also look for indications about the building height or volume to later classify residential houses in single family or multi-family houses. Only 5% of the houses in Germany exhibit a direct tag containing this information.

- *Points and Regions of Interest.* Besides buildings, there are other indicators for low or high population numbers. Specifically, we consider the boundaries for parks, playgrounds and tourist attractions of any kind.

- *Landuse and Places.* In OSM, there is a designated tag to assign landuse to an area (typically described via an OSM relation). Examples are residential area, commercial area, industrial area, forest, meadow, or farmyard. The majority of buildings in Germany (91.5%) are located in an area with specified landuse. Moreover, we consider place categories (villages, towns, suburbs, etc.).

- *Street Network Data.* A street is typically stored as a way in OSM. While the street names are not important for our application, the street categories (highway, living road, pedestrian, etc.) are. The street network is used in the baseline approach (based on a Voronoi diagram) as well as in our approach.

Finally, we also extract the boundaries of selected cities and villages for which we know the precise census data from OpenGeoDB. This information will be used in the very last step of our framework. Note, that in contrast to most previous work, we only use census data on a coarse-grained level, and only in the training phase, not when predicting.

3. ENRICHING OSM DATA

As discussed above, building tags necessary for precise population estimation are incomplete in OSM. We now describe how to learn the missing information from available data. This step in our approach could hence also be used as a standalone automatic tool for enhancing the OSM data coverage.

For other parts of the data, like street categories, similar approaches were investigated before [9]. To the best of our knowledge, we are the first to present a learning-based approach that predicts new landuse and building tags. In order to compute the required features efficiently, we store all extracted entities in an R-Tree [3]. For ways or relations, we store the respective centroid.

3.1 Landuse Classification

In a later step, we want to be able to classify buildings as residential vs. non-residential. We therefore first like to know the landuse of the area the building is located in. A house without a tag indicating that it is a business of any kind in a residential area is very likely to be a living house,

[4] `http:www.opengeodb.org`

[5] `http://wiki.openstreetmap.org/wiki/MapFeatures`

whereas a building in an industrial area is not. As mentioned before, only about a tenth of all buildings in Germany are not within an area of known landuse. These are perfect conditions for a learning-based approach, since there is a large and diverse set of training data.

We first subdivide areas without a landuse tags in smaller units, using large streets as boundaries. As long as the area of the unit does not drop below a certain threshold (50,000 square meters), we determine the most important street in the area (using OSM street categories) and use it to cut the unit in half (possibly elongating the street to obtain a complete cut). We here assume that the type of landuse is consistent within a sufficiently small area.

We use the following features in our approach to classify the computed units: contained or near-by *place specifications* such as hamlet or suburbs, the *building density* expressed by the total number of buildings in the area, the average *building size*, the *number of schools and universities* in with a distance of at most three kilometers from the unit, and the *number of leisure facilities, playgrounds and parks* within a distance of 500 meters from the unit. For each of these features, a high number is indicative of a residential area.

Furthermore, we consider the *number of small shops*, the *number of offices* and the *number of craft producers*. For each of these features, a high number makes a residential area less likely. Moreover, we consider the *street density* for living streets and others. A high number of living streets makes a residential area more likely, a high number of streets of other types less likely. We query a pre-computed R-Tree to obtain all these numbers.

3.2 Inhabited or Uninhabited Buildings

Many existing approaches for population estimation distribute the people living in an area uniformly among the buildings in that area or proportional to the building sizes. This leads to distortions when the area also contains industrial and commercial buildings, or hospitals, schools, police stations, garages and the like. We therefore try to identify uninhabited buildings so that we can ignore them for the final population estimation.

The OSM tags only classify a small fraction of all buildings into inhabited and uninhabited. Figure 3 shows examples of industry buildings which are not tagged as such. Our goal is to, again, learn the correct values for the missing tags. One important feature is the *landuse*. As explained above, buildings in residential areas without any tag indicating a business are very likely to be inhabited. Additionally, we consider the *base area* of the building. The intuition is that very large base areas are often indicative for factories or supermarkets. We also regard the *complexity of the base area* by counting the number of nodes on the OSM way that describes the building footprint. Moreover, we use all the features described above for the landuse classification but with a smaller distance from the object to be classified (which are now individual buildings instead of areas).

The training data is as follows. For uninhabited buildings, we use the 6% of buildings in Germany with an amenity tag that clearly marks them as non-residential. For inhabited building, we choose houses with a suitable tag as well as houses located in well-mapped residential areas without any tag.

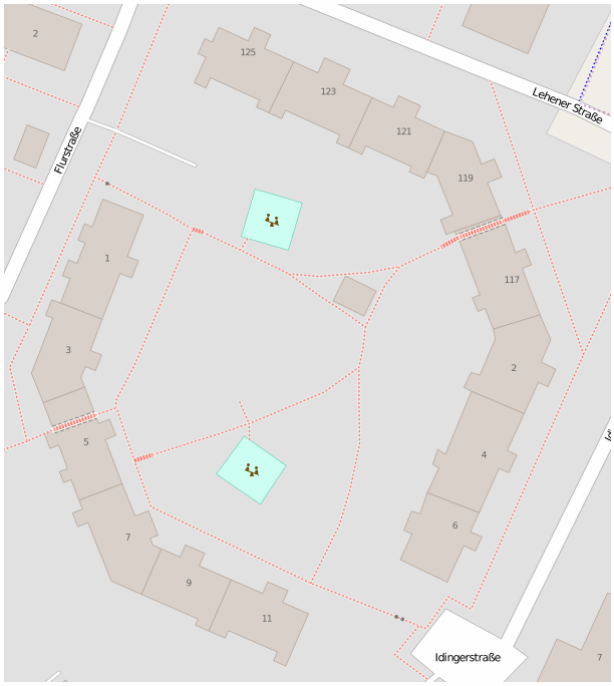

Figure 4: As there are two playgrounds in the inner courtyard, the surrounding buildings are likely to accommodate multiple families each despite their rather small base areas.

3.3 Single or Multi-Family Houses

The number of people living in a single building naturally varies to a great extent. In particular, the number of inhabitants in multi-family houses and apartment buildings is usually significantly higher than in single family houses. The OSM data is very incomplete when it comes to building heights and tags that indicate how many floors or flats are in a building. Our goal is again to learn the missing information.

If building footprints are available, the *base area* is again a good feature for classification. Also the *number of entrances* or *address specifications* provides some insight. Moreover, we again use the *number of shops, parks and playgrounds* in the neighborhood together with the *building density* as features. For instance, if there are many playgrounds but few buildings, these buildings are likely to accommodate many families; see Figure 4 for an example. Another feature is the *type or place* of the locality: we differentiate between hamlets, villages, suburbs and towns. The idea is that villages and hamlets tend to have less multi-story buildings.

Our training data is as follows. In the OSM data, about 6% of the buildings have a tag that mark them as single-family houses and 0.5% of the buildings are marked as apartments (multi-story building).

3.4 Classifier Choice

We expect the relationship between many of the features and the target to be rather simple. For example, we expect that the larger the number of factories the more likely it is that an area's landuse is industrial. For such feature-target correlations, a suitable learning method is logistic regression.

Nevertheless, we also have to deal with some more com-

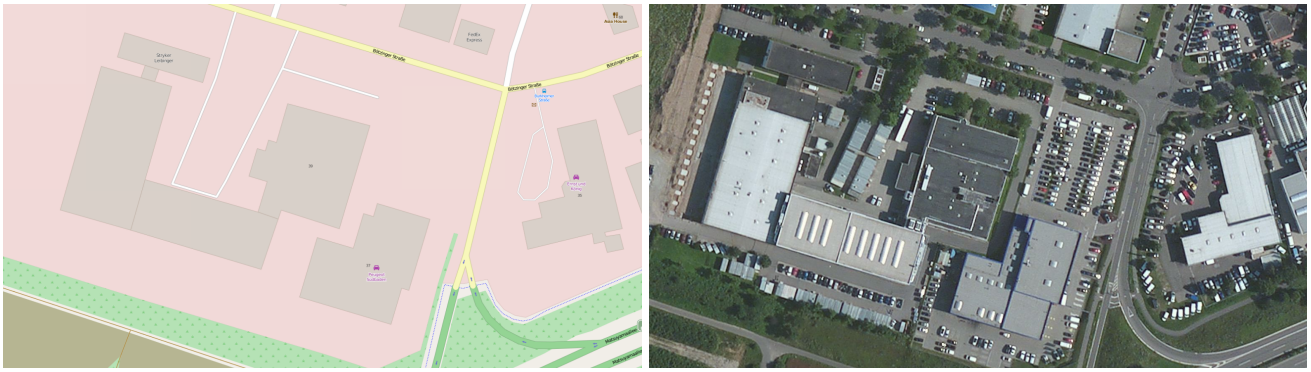

Figure 3: The left image depicts an industrial area according to OSM data. Several of the buildings inside do not exhibit amenity or name tags. Still these are no living houses as to be observed in the satellite image on the right. But being located in an industrial zone and having a large base area allows to classify these buildings correctly as uninhabited with our approach.

plex relationships between features and targets. For example, a very large base area makes a building unlikely to be a residential building. But so does a very small area, since it is indicative of something like a garage or shed and not an inhabited house. So for this feature, the correlation to the target is not linear but reaches its maximum for medium values.

Also some features need to be considered in combination. For example, an untagged building in a residential area is very likely to be inhabited, whereas an untagged building in a non-residential area is not. For such conditional (if ... then ... else) dependencies, random forests are the method of choice. Both of these classifiers often work well with default parameters (using, e.g., scikit-learn [14]) and the learned models are easy to interpret.

4. POPULATION ESTIMATION

At this point we are able to differentiate between different types of areas and buildings. This gives us insight into the urban structures of a city. We can exclude sets of buildings from being populated at all, which enhances the accuracy of our estimate primarily in commercial and industrial areas. We know whether a building is a multi-family house or a single-family house. Combined with the data about the locality we obtain information about the specific character of an area. For example, many multi-family houses within a city's center suggest a congested urban area. With all those features we are finally able to distribute a given population among the buildings of a certain area. To be independent from local census data we extrapolate the results we gain for a small number of cities and villages, again with the help of machine learning.

4.1 Feature Selection

There are basically two kinds of information we are trying to incorporate in the extrapolation: information about the building itself and information about its environment.

To describe the building itself we use its *base area*, if its a single-family house or a multi-family house and if the building contains a business of any kind.

To describe the environment we use previously obtained *landuse* and *place* data because buildings within villages and suburbs are not as densely populated as those within a cities

Figure 5: Houses with similar base area in a city. Houses on the left side of the large street are likely to have less floors and accommodate less occupants than houses on the right, though.

core. To provide a measure of the size and populousness of the environment we consider the number of *schools*, *shopping malls* within a distance of up to 5 kilometers, the number of *suburbs*, *villages*, *cities* within a distance of up to 30 kilometers, and the size of the city where the building is located. Finally, we also use the number of *shops*, *parks*, and *playgrounds* within the neighborhood.

Figure 5 shows that both aspects, the building itself and its environment, are important for correct population estimation.

4.2 Obtaining Training Data

We take cities, city districts and villages of various sizes and distribute respective census data from OpenGeoDB among local buildings with a weighted areal interpolation scheme [17] to provide a ground truth for the following extrapolation.

$$\text{population}(i) = \frac{\text{weight}(i) \cdot \text{area}(i)}{\sum_{b \in \text{buildings}} \text{weight}(b) \cdot \text{area}(b)} \cdot \text{total\_population}$$

To determine the weight we use multiple parameters. Uninhabited houses have a weight of zero. For inhabited houses, one parameter is the type of building (single-family vs. multi-

Figure 6: Small village in OSM. The grey area is tagged with *landuse=residential* and the streets also indicate that people are living in that area, but no buildings are present at all.

family). A building receives a penalty if the building is not only a living house but also contains a shop or some kind of business. A penalty is also given for buildings within hamlets and villages, because they are not as densely populated as in towns. Another parameter is the landuse of the containing area.

4.3 Machine Learning Approach

One method to learn the number of inhabitants per building is linear regression. As explained above, this makes sense for linear feature-target correlations: for example, a larger base area usually means more inhabitants. Besides linear regression, we also use regression trees [11]: these are a variant of decision trees which are suitable to learn real-valued functions instead of classifiers.

4.4 Handling Areas without Buildings

In [4], the number of buildings in Germany is estimated to be about 38 million. The current number of building footprints in the OSM data is about 18.9 millions. If individual buildings are missing in otherwise well-mapped areas, the distortion in the population estimation is minor. But if whole residential areas are without any buildings (see Figure 6 for an example), our approach as described so far fails.

We found this to happen almost exclusively for small villages. We then make use of two other clues about the population number: the area of the village and the street network data. We could use the described Voronoi-baseline approach as a fall-back since it only relies on street network data. But we decided to learn a population estimator purely based on the area and the accumulated street length. Due to the missing building footprints, we use simple areal interpolation by distributing the inhabitants uniformly in the village area.

5. EXPERIMENTAL EVALUATION

In this section, we first evaluate each of the described classifiers to enrich the OSM data. Subsequently, we show that the selected features for population estimation allow for accurate predictions in metropolitan as well as rural areas.

Target	Prediction			Precision
	resid.	indus.	comm.	
residential	841	83	76	84.10%
industrial	89	778	133	77.80%
commercial	111	212	677	67.70%
Recall	80.79%	72.51%	76.41%	76.53%

Table 1: Results for landuse classification with our approach, based on 1000 samples for each class.

5.1 Data and Settings

We downloaded the OSM data for Germany in XML format (16-02-2015)[6] and extracted the relevant parts for our application as listed in Section 2. Our experiments were conducted on an AMD FX(tm)-8150 Processor with 3.6 GHz and 32 GB RAM. Moreover, we downloaded boundary data of various European and other cities[7] and subsequently extracted the same kind of data as we did for Germany.

5.2 Classifier Evaluation

We first provide results on Baden-Württemberg, a large state in Germany with an area of 35,751.5 km^2 and about 10.7 million inhabitants. We present detailed results on the quality of our three learned classifiers that we later use for the final population estimation: kind of landuse, inhabited vs. uninhabited buildings, and single-family vs. multi-family houses. For every class we used a set of 1,000 randomly chosen samples from the existing training data in the learning phase. We also conducted experiments with larger samples, but observed no significant difference in the quality of the results.

5.2.1 *Landuse Classification*

The first step in our framework is the landuse classification. We distinguish between residential, commercial and industrial landuse. Figure 1 shows the prediction accuracy of our random forests classifier. For residential areas, we correctly predict the landuse in over 84% of all cases. The results for commercial and industrial areas are slightly worse. The main source of misclassification is when there are too few buildings with any tag at all in the area of interest or close-by. We observe that a mix-up between residential and commercial landuse is more likely than a mix-up between residential and industrial landuse. The reason is that, for the latter, features that are not based on building tags (e.g., building and street density) allow for a better distinction. Classification with logistic regression instead of with random forests leads to an overall accuracy of only 63%. This shows that the complex relationship between our features is better captured using random forests.

We used our classifier to predict the landuse for 8.5% of uncovered buildings in OSM. Most of them were classified as residential. Manual sample checks using satellite images showed that the classifier works well and therefore allows to enrich the OSM data.

5.2.2 *Inhabited vs. Uninhabited Buildings*

The next step is to decide whether a building has inhabitants or not.

[6] `www.geofabrik.de`
[7] `https://osm.wno-edv-service.de/boundaries/`

Target	Prediction		Precision
	inhab.	uninhab.	
inhabited	976	24	97.60%
uninhabited	72	928	92.80%
Recall	93.13%	97.48%	95.20%

Table 2: Quality analysis of our learned classifier for inhabited vs. uninhabited buildings, based on 1000 samples for each class.

Target	Prediction		Precision
	multi	single	
multi	927	73	92.70%
single	108	892	89.20%
Recall	89.57%	92.44%	90.95%

Table 3: Prediction quality for single-family vs. multi-family houses, based on 1000 samples for each class.

We computed *area under curve (AUC)* scores for all our selected features. We observe highest correlations for base area size, residential landuse, number of buildings in a radius of 25 and 50 meters and number of garages in a radius of 25 to 500 meters. Figure 2 shows that a random forests classifier gives very accurate predictions, with an overall precision of 95.2%. Again, classification using logistic regression performed worse with an overall accuracy of 82.8%.

5.2.3 Single-Family vs. Multi-Family Houses

The last step of supplementing the OSM data consists of dividing the residential buildings into single-family and multi-family houses. We rely on the previous learning step here and only consider buildings which we classified as inhabited. Figure 7 illustrates the learned feature weights using Random Forest. We see that closeness to several facilities and parking lots plays an important role in the classification process, as well as base-area size and complexity (as measured by the number of nodes in the footprint).

As shown in Figure 3, the prediction accuracy is not as good as for the inhabited vs. uninhabited classification step. But still over 90% of buildings are classified correctly.

5.3 Population Number Estimates

Finally, we want to evaluate how accurate the population estimation based on existing and newly learned OSM data is. Since we do not have access to fine-grained census data, we cannot compute the accuracy of our approach on the level of individual buildings. Instead, we aggregate the number of inhabitants in different kinds of regions and compare these numbers to the census data for the respective region, as obtained from OpenGeoDB.

We used three benchmark sets to evaluate the population estimates for Baden-Würtemberg. The first set contains large cities, the second set contains medium-sized cities and towns, and the third set contains villages and rural areas. The places were chosen randomly for each set. For each place, we computed the population numbers using the Voronoi-baseline as well as our approach.

We learned feature weights on a set of 20 selected districts, villages, towns and cities after interpolating the population numbers for these regions as described in Section 4.2. Then we computed the respective feature vector for every building in Baden-Württemberg and estimated the number of inhabitants on that basis. To compare our results to coarse-grained census data, we aggregated the population numbers of buildings in our benchmark regions. Table 4 shows an excerpt of the results. We make the following observations:

- The population numbers for large cities are fairly accurate when using our estimation tool. On average, the number differs less than 10% from census data. For Mannheim, Karlsruhe and Heilbronn our estimate matches the value from the census data almost perfectly. The estimates of the baseline approach are far too small, because the accumulated length and density of the streets does not reveal the population density along those streets sufficiently.

- For villages and small towns, the estimation quality is not as good as for cities, with an average difference of about 25% from census data. The reason is that here small absolute errors translate into large relative errors. In a village with 1,000 similar living houses it is very difficult to determine for each of these houses, whether it is inhabited by a single person, a couple or a family with one or more children. This remains true even if complete information about the size and nature of the houses would be available. This could easily lead to a factor of two or more of over- or underestimation. Therefore, we deem our results with a maximal difference of 53% to be of good quality also for this benchmark set.

  Figure 8 depicts the region with the maximal relative error (of our tool as well as of the baseline approach). It is a rather large region with long streets but only very few houses and inhabitants. This is particularly problematic for the baseline approach, which overestimates the population in this region by a factor of more than 10.

- For medium-sized cities, the average deviation from the census data is about 18%. The deviation depends strongly on the characteristics and the structure of the city. For example, for cities consisting of a main part and several incorporated villages the estimation was worse than for more homogeneous cities.

Our approach outperforms the baseline in over 90% of the tested benchmark regions. Moreover, the maximal absolute and relative error are both much smaller with our approach.

5.4 Scalability to Germany and Other Countries

Next, we want to show that the classifiers learned on Baden-Württemberg data are suitable to compute population numbers for other parts of Germany as well. For that purpose, we selected seven cities in Germany and compared the census data to our estimation. The results are provided in Table 5. We observe that for all cities but Oldenburg the estimated number of inhabitants is quite accurate. For Oldenburg we predicted almost twice the number of real inhabitants. This can be explained by a very dense building development in the city center as illustrated in Figure 9. There are plenty of buildings without a road connection and

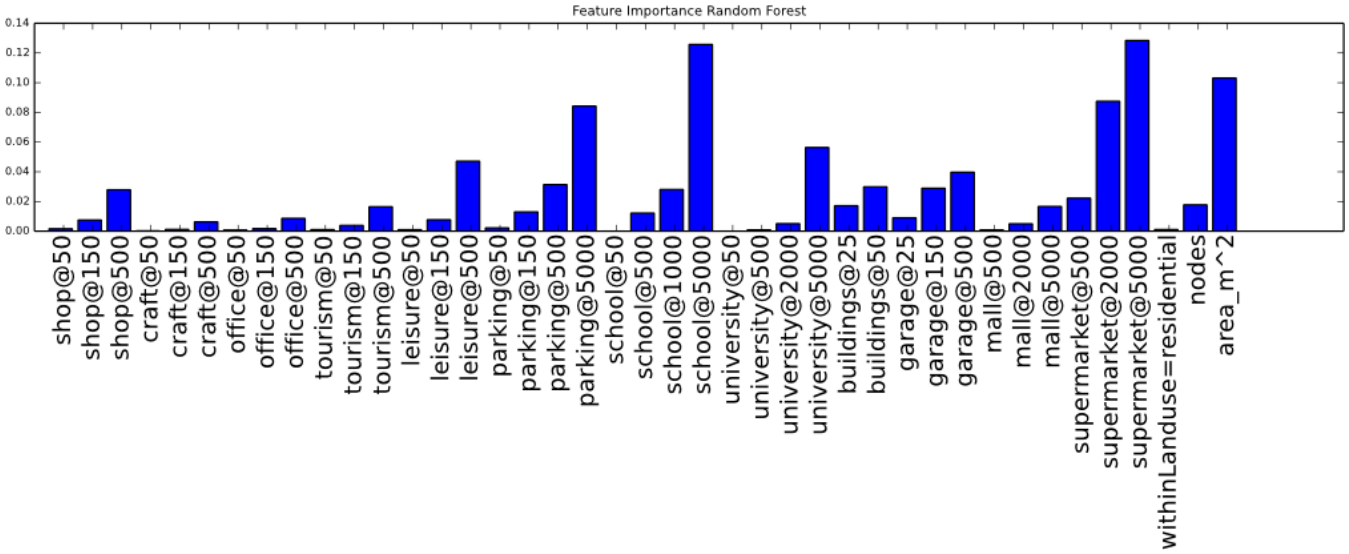

Figure 7: Learned feature weights for single/multi-house classification using random forests.

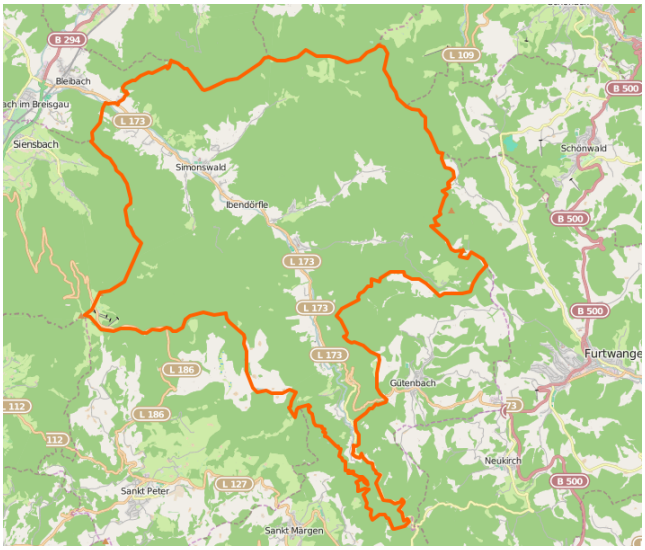

Figure 8: Benchmark region with maximum relative error for the baseline (1085%) as well as for our approach (53%).

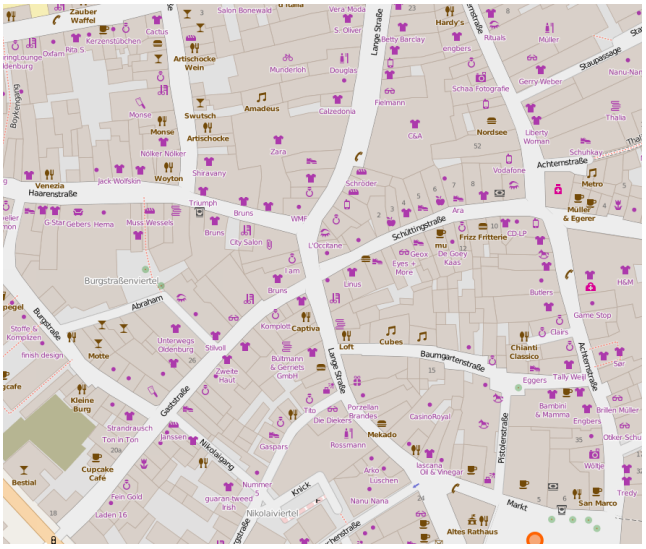

Figure 9: Cut-out of the city center of Oldenburg.

a large number of shops of all kinds. This leads our tool to overestimate the building heights and number of people living in that area.

For the other cities we are on average only 10.5% away from the census data. This is consistent with our results on large cities in Baden-Württemberg.

To put the numbers into perspective, one has to consider that, over time, the population numbers vary significantly, in particular in large cities. For example, Hamburg had 1.707 million inhabitants in 2011, which increased to 1.799 million in 2012. This is a plus of about 5%. In 2013, the number dropped again to 1.734 million. Of course, also the number and shape of buildings in the city changes, but population numbers are prone to change more quickly.

Finally, we went one step further and also considered cities in other countries. We checked six large European cities as well as one city in the U.S. of A. (Seattle). Table 6 shows that, even for these cities, our classifiers (trained on parts of Germany) produces useful estimates. Of course, we could learn our classifiers individually for each country, or even on a more fine-grained level, and thus get even better results. But the numbers in Table 6 indicate that our approach extrapolates surprisingly well (with an average relative error of only 12.5%) and that the selected features seem to be valid in other countries, too.

Region		Population	Abs. Error	Rel. Error
Burgholzhof	C	2774		
	B	908	-1866	-67%
	E	2156	-618	-22%
Gundelfingen	C	11554		
	B	12710	1156	10%
	E	11124	-430	-4%
Heitersheim	C	5968		
	B	12452	6484	109%
	E	3298	-2670	-45%
Kirchzarten	C	9758		
	B	14405	4647	48%
	E	6836	-2922	-30%
Kenzingen	C	9518		
	B	31194	21676	228%
	E	12284	2766	29%
Opfingen	C	4108		
	B	13989	9881	241%
	E	3475	-633	-15%
Ihringen	C	5865		
	B	23366	17501	298%
	E	6241	376	6%
Simonswald	C	3024		
	B	35838	32814	1085%
	E	4632	1608	53%
Herbolzheim	C	10251		
	B	34642	24391	238%
	E	14916	4665	46%
Stühlinger	C	18316		
	B	5917	-12399	-68%
	E	16791	-1525	-8%
Müllheim	C	18454		
	B	49096	30642	166%
	E	15752	-2702	-15%
Fellbach	C	44403		
	B	28856	-15547	-35%
	E	33152	-11251	-25%
Lörrach	C	48307		
	B	34216	-14091	-29%
	E	44026	-4281	-9%
Baden-Baden	C	53012		
	B	103036	50024	94%
	E	42905	-10107	-19%
Ulm	C	119218		
	B	127389	8171	7%
	E	101931	-17287	-15%
Stuttgart	C	604297		
	B	294367	-309930	-51%
	E	538410	-65887	-11%
Pforzheim	C	117754		
	B	101136	-16618	-14%
	E	83251	-34503	-29%
Mannheim	C	296690		
	B	203605	-93085	-31%
	E	296930	240	0%
Karlsruhe	C	299103		
	B	216051	-83052	-28%
	E	297266	-1837	-1%
Heilbronn	C	118122		
	B	116572	-1550	-1%
	E	116905	-1217	-1%
Heidelberg	C	152113		
	B	119242	-32871	-22%
	E	166390	14277	9%
Freiburg im Breisgau	C	220286		
	B	172285	-48001	-22%
	E	226803	6517	3%

Table 4: Results for population estimation: B denotes the Voronoi-baseline, E our estimation and C the census data.

City		Population	Abs. Error	Rel. Error
Munich	C	1388852		
	E	990396	-398456	-29%
Aachen	C	240086		
	E	235124	-4962	-2%
Hannover	C	514137		
	E	573738	59601	11%
Münster	C	296599		
	E	314471	17872	6%
Düsseldorf	C	593682		
	E	657101	63419	10%
Hamburg	C	1734420		
	E	1639908	-94512	-5%
Oldenburg	C	159610		
	E	301699	142089	89%

Table 5: Population estimation for German cities. C denotes census data, E the estimated value.

City		Population	Abs. Error	Rel. Error
Marselle	C	850726		
	E	606168	-244558	-28%
Brussels	C	1138854		
	E	1099289	-39565	-3%
Bern	C	126598		
	E	129903	3305	3%
Copenhagen	C	583348		
	E	465761	-117587	-20%
Cambridge	C	123867		
	E	143693	19826	16%
Seattle	C	652405		
	E	684084	31679	5%

Table 6: Population estimation for European and U.S. cities.

6. CONCLUSIONS AND FUTURE WORK

We presented a framework for population estimation which uses OpenStreetMap data to estimate population counts on the level of individual buildings. Our approach uses census data and areal interpolation only to create training data. From these, we can *learn* population numbers for areas without any available census data. Therefore, in contrast to most previous work, our framework does not only allow for interpolation but also for extrapolation. Our experiments showed the potential of our approach to estimate population numbers world-wide, while using only a very limited amount of training data.

There are various possibilities to improve our framework. First, better OpenStreetMap data will eventually lead to better population estimates automatically. For example, a better coverage of building footprints or amenity tags would be helpful. Moreover, the estimation quality for medium-sized cities could possibly be improved when regarding additional features. Often they are composed of the actual city and several incorporated villages. A partitioning approach which recognizes areas of homogeneous structure could be used to take care of such differences. Historical data could contribute here as well. Other additional information sources, such as traffic flows (from openly available GPS tracks), could further improve the quality of our estimates.

7. REFERENCES

[1] Mohamed Bakillah, Steve Liang, Amin Mobasheri, Jamal Jokar Arsanjani, and Alexander Zipf. Fine-resolution population mapping using openstreetmap points-of-interest. *International Journal of Geographical Information Science*, 28(9):1940–1963, 2014.

[2] Hannah Bast, Jonas Sternisko, and Sabine Storandt. Forestmaps: A computational model and visualization for forest utilization. In *Web and Wireless Geographical Information Systems*, pages 115–133. Springer, 2014.

[3] Norbert Beckmann, Hans-Peter Kriegel, Ralf Schneider, and Bernhard Seeger. The r*-tree: an efficient and robust access method for points and rectangles. In *SIGMOD*, 1990.

[4] Martin Behnisch and Alfred Ultsch. Estimating the number of buildings in germany. In *Advances in Data Analysis, Data Handling and Business Intelligence*, pages 585–593. Springer, 2010.

[5] Robert G Cromley, Dean M Hanink, and George C Bentley. A quantile regression approach to areal interpolation. *Annals of the Association of American Geographers*, 102(4):763–777, 2012.

[6] Jerome E Dobson, Edward A Bright, Phillip R Coleman, Richard C Durfee, and Brian A Worley. Landscan: a global population database for estimating populations at risk. *Photogrammetric engineering and remote sensing*, 66(7):849–857, 2000.

[7] Cory L Eicher and Cynthia A Brewer. Dasymetric mapping and areal interpolation: Implementation and evaluation. *Cartography and Geographic Information Science*, 28(2):125–138, 2001.

[8] Robert Hecht, Carola Kunze, and Stefan Hahmann. Measuring completeness of building footprints in openstreetmap over space and time. *ISPRS International Journal of Geo-Information*, 2(4):1066–1091, 2013.

[9] Musfira Jilani, Padraig Corcoran, and Michela Bertolotto. Automated highway tag assessment of openstreetmap road networks. In *SIGSPATIAL*, 2014.

[10] Kate E Jones, Nikkita G Patel, Marc A Levy, Adam Storeygard, Deborah Balk, John L Gittleman, and Peter Daszak. Global trends in emerging infectious diseases. *Nature*, 451(7181):990–993, 2008.

[11] Roger J Lewis. An introduction to classification and regression tree (cart) analysis. In *Annual Meeting of the Society for Academic Emergency Medicine in San Francisco, California*, pages 1–14, 2000.

[12] Zhenyu Lu, Jungho Im, Lindi Quackenbush, and Kerry Halligan. Population estimation based on multi-sensor data fusion. *International Journal of Remote Sensing*, 31(21):5587–5604, 2010.

[13] K Rask Overgaard. Urban transportation planning'traffic estimation. *Traffic Quarterly*, 21(2), 1967.

[14] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] Francesca Pozzi, Christopher Small, and Gregory Yetman. Modeling the distribution of human population with nighttime satellite imagery and gridded population of the world. *Earth Observation Magazine*, 12(4):24–30, 2003.

[16] Anna F Tapp. Areal interpolation and dasymetric mapping methods using local ancillary data sources. *Cartography and Geographic Information Science*, 37(3):215–228, 2010.

[17] Serkan Ural, Ejaz Hussain, and Jie Shan. Building population mapping with aerial imagery and gis data. *International Journal of Applied Earth Observation and Geoinformation*, 13(6):841–852, 2011.

[18] Takeo Yamada. A network flow approach to a city emergency evacuation planning. *International Journal of Systems Science*, 27(10):931–936, 1996.

[19] Paul A Zandbergen and Drew A Ignizio. Comparison of dasymetric mapping techniques for small-area population estimates. *Cartography and Geographic Information Science*, 37(3):199–214, 2010.

[20] Caiyun Zhang and Fang Qiu. A point-based intelligent approach to areal interpolation. *The Professional Geographer*, 63(2):262–276, 2011.