Poverty Distribution Mapping Using a Random Forest Algorithm

A Case Study of the Mekong River Delta

Word Count: 6.315

Lars Kouwenhoven (AUC)
Science Major
larskouwenhoven@me.com

Supervisor:
Eric Koomen (VU)
e.koomen@vu.nl

Reader:
Maurice de Kleijn (VU)
mtm.de.kleijn@vu.nl

Tutor:
Michael McAssey (AUC)
m.p.mcassey@auc.nl

Abstract

In order to tackle some of the challenges the poor face, it is vital to have data on how they are distributed. However, this data is often lacking exactly in countries and regions where a large portion of the population is considered poor. The aim of this research is to explore whether a Random Forest algorithm can be employed to produce low-level poverty data, based on high-level poverty information, focusing on the rural Vietnamese Mekong river delta region. This region is of interest since it faces potentially increased natural hazard risks, of which it is expected that those in poverty are affected strongly. Then, seeing that the Mekong river delta is one of the poorer areas in Vietnam, with large socioeconomic inequalities, it is vital to accurately map the distribution of the poor in order to find remedies to the challenges they face. A method of applying machine learning to a combination of point of interest data and ancillary geospatial information is proposed here, with the aim of mapping local poverty levels on a 5 spatial kilometer resolution. Weaknesses and strengths of such a method are discussed, and compared to results of similar research performed previously. It is concluded that this research gives an indication that this method is indeed successful in the Mekong River Delta, although additional low-level validation data is required to quantify this statement. Additionally, this method may be applied in other global regions with sufficient availability of base data, where similar efforts will increase the likelihood of the poor receiving aid effectively, be it from local governments, NGOs or other organizations.

Keywords: Random Forest, Mekong River Delta, point of interest, poverty distribution

1. Introduction

It is an unfortunate reality that not all individuals are exposed to the same risks caused by pollution and natural hazards. Specifically, the poor have been shown to be disproportionately exposed to these factors (Winsemius et al., 2018). In order to mitigate these effects for the poor, however, it is vital to have knowledge about how they are spatially distributed. Problematically, it is exactly in countries with a large part of the population living in poverty, where data on wealth distribution is lacking.

The Mekong river delta is a particularly interesting area of study. Due to its geographical features, it is vulnerable to natural hazards (Hoang et al., 2018). Additionally, there is a relatively big income inequality, that, unlike the case in some of the other regions in Vietnam, has improved only slightly during the past ten years (World Bank, 2018). Worse still, in the period between 2014 and 2016 the Gini coefficient, a measure of inequality, even rose 2 percentage points. Lastly, partially due to the rural nature of the region, data on local wealth distribution is limited (Bali et al., 2008).

Using machine learning to predict local wealth distribution based on satellite imagery is a relatively new field of research. Until some 5 years ago, the algorithms required were not yet powerful or manageable enough to process large sets of data (Njuguna and McSharry, 2017). However, more research has been performed on a comparable challenge, with the goal of mapping population distributions using machine learning and satellite imagery (Stevens et al., 2015; Gaughen et al., 2016; Ye et al., 2018). This research commonly employs a random forest (RF) algorithm, using remote sensing (satellite) data and other geospatial data (roads, waterdays, elevation). The results are promising, though it is noted that results are hard to verify, and where they are verifiable, it is unknown whether conclusions can be extended to the entire area of research (Gaughen et al., 2016). Ye et al. (2018) find excellent results through the use of point of interest data from Baidu, a major Chinese map maker, with assumed near-complete coverage of China. Amongst the authors, there is an ongoing discussion around the value of night-time light (NTL) data: while Ye et al. (2018) argue that

is has little value due to blooming, the other authors successfully employed it in their algorithms. As opposed to a random forest algorithm, a convolutional neural network (CNN) has been applied by some to estimate poverty distribution (Jean et al., 2016; Perez et al., 2017). CNNs truly constitute a 'black box': it is entirely impossible for humans to interpret the way in which a CNN is trained and comes to its prediction. While it may be worthwhile to apply a CNN when the robustness of predictions using a RF are verified, the research surrounding poverty prediction in the Mekong River Delta has not reached such a stage, presently.

This research provides a method of generating poverty level data, applying a Random Forest machine learning algorithm to publicly available ancillary geospatial and point-of-information data. It extends upon previous research by using point of information (POI) data in addition to geospatial and remote sensing data, applied to the Vietnam Mekong River Delta. In doing so, the proposed research will contribute to the ongoing NWO project Adaptation Pathways for socially inclusive development of urbanizing deltas. Concretely, this research attempts to answer the central question: to what extent is a Random Forest model capable of producing local poverty data based on poverty data on a higher level, supplied by ancillary spatial data? This will be done by answering this question applied to the Vietnamese Mekong River Delta. Then, while minding the context and specificalities of this region, a general answer can be formed. If this method performs well in the Mekong river delta, it may be extended to other locations globally, with the promise of accurately helping those in need of support due to pollution and natural hazards, amongst a wide variety of applications.

2. Materials and methods

In order to find answers to the questions posed in this research, a step-wise approach will be taken. This process is displayed schematically in *Figure 1*, and will be detailed in this section.

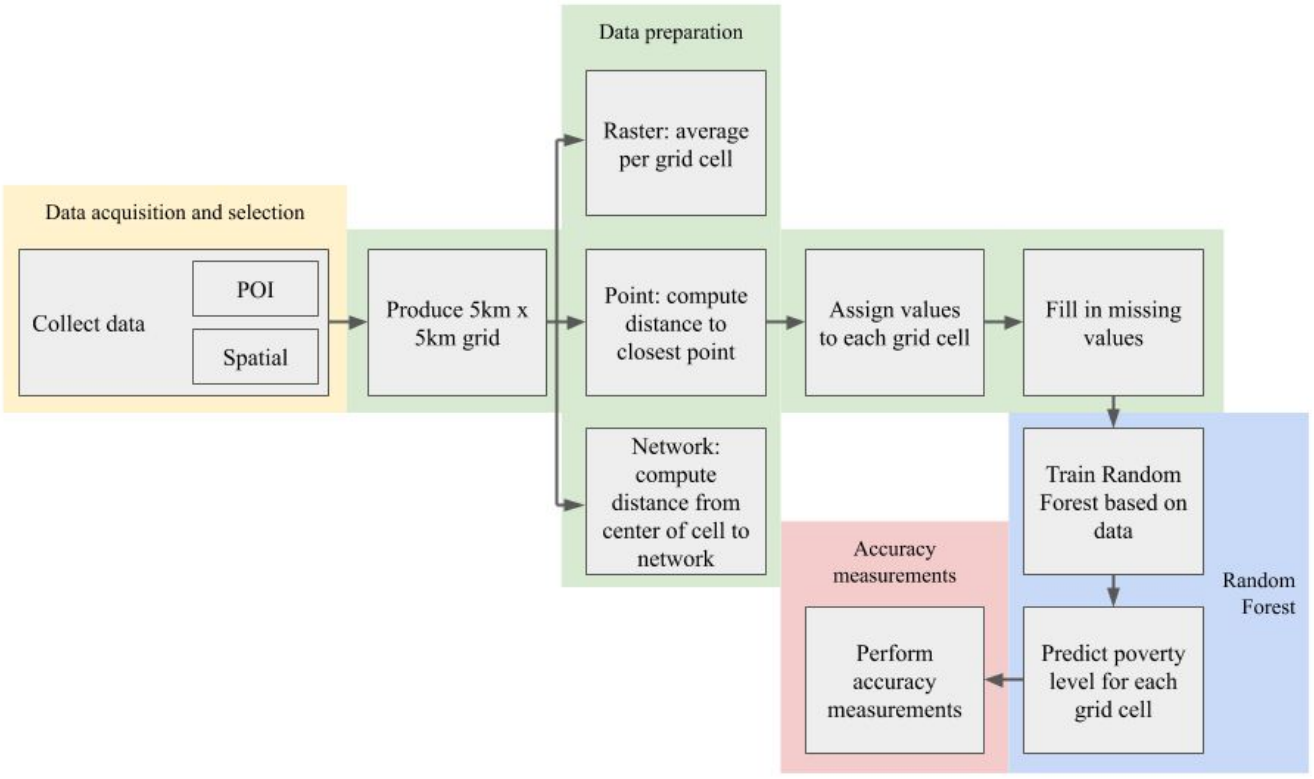

Figure 1: Schematic overview of processes described in section 2

2.1 Data acquisition and selection

Vital to the performance of the Random Forest algorithm, as with any form of machine learning, is the preselection of data to be used (Breiman, 2001). In the field of poverty and population mapping, two forms of data are generally used: remote sensing (e.g. satellite imagery, land cover, nighttime light) data and geospatial (e.g. road networks, population densities, built up area) data (Stevens et al., 2015; Gaughen et al., 2016; Ye et al., 2018). Recently, the use point of interest data has demonstrated promising results, indicating that it is a valuable asset in population or poverty

mapping (Ye et al., 2018). A POI is an item of particular interest to some individual, on a map represented by a point. Examples of POIs include schools, hospitals and government buildings. They are distinguished from other items displayed on a map, in the sense that they are best represented by a singular point, as opposed to a linear feature, such as a railway or river.

In general, data is selected under two conditions: (1) availability and (2) expectation of predictive value (Stevens et al., 2015; Gaughen et al., 2016). The first condition is not always so much a form of selection, but rather using all resources that are available. It is a frequent occurrence that data one may hope to use is in fact unavailable, be it due to licensing and privacy issues, or because the data is simply non-existent at this point it time. Although it may be regarded a valid method to select data based on its predictive power, the second condition, this may lead to the exclusion of factors that unexpectedly contribute to the variable one is attempting to map. Additionally, as described in section 2.3 in greater detail, the Random Forest algorithm has been demonstrated to be unsusceptible to irrelevant data (Wiener and Liaw, 2002). As such, there is no harm in adding data that is expected to be irrelevant.

While keeping these considerations in mind, a selection of data to be used was then made for this research, displayed in Table 1. First of all, in the category remote sensing data, only land use data from MODIS was selected. Even though nighttime light data has been employed by Stevens et al. (2015) and Ye et al. (2018), Mellander et al. (2015) have demonstrated it contains little capacity when applied to the mapping of economic activity. As a result, nighttime light data was not employed in this research. Additionally, satellite data was not used directly. Since in the basis, satellite data consists of RGB values, it was deemed to contain little explanatory power for the dependent variable 'poverty level'. Additionally, it must be noted that satellite data was used *indirectly*, since the land use data published by MODIS is based on satellite imagery (Channan, Collins, and Emanuel, 2014).

Following, a series of geospatial data has been selected for use in this research. These datasets were predominantly chosen due to the condition 'availability': they are amongst a select few sets of

data focusing on the Mekong River Delta that are publicly accessible. An overview of the sets can be found in *Table 1*.

Type	Variable Name	Description	Data Source	Resolution	Year
Raster	landcover	12 categories of land cover in the Mekong River Delta	MODIS	500m	2014
	Recur_XX	Flood recurrence per province			2018
	GHS_builtup	Built area per province	European Commission	38m	1975 1990 2000 2014
	GHS_pop	Population density per province	European Commission	38m	
Vector	road_street	Roads and streets in the delta	WISDOM		
	road_srteet_new	New roads	WISDOM		
	national_highways	National highways	WISDOM		
	provincial_roads	Provincial roads	WISDOM		
	residential_roads	Residential roads	WISDOM		
	CanalRiver	Rivers and canals	WISDOM		

Table 1: Forms of data used

In addition to the two traditional types of data used in similar research, point of interest data was used for its promising recent results. Although Ye et al. (2018) made use of data from Baidu Map Services, which provides a (near-) complete set of Chinese point of interest data, no such complete openly accessible data exists for Vietnam, or the Mekong River Delta specifically. Therefore, OpenStreetMap (OSM) was used as the source of POI data for this research. OSM is one of the earliest and most popular examples of volunteered geographic information (VGI), allowing any internet user around the world to add information related to localities they are familiar with (Mooni and Minghini, 2017). Since data is added by volunteers and not always verified, there is an ongoing debate regarding the quality of data provided by OpenStreetMap (Barron, Neis & Zipf, 2014). Quality can mainly be compromised in the following manners: (1) completeness of the data set, (2) positional

accuracy of items of the data set and (3) the accuracy of categories attributed to items of the data set (Barron, Neis & Zipf, 2014). Comparing OSM data with a verified reference data set, for example provided by a government, constitutes the most straightforward and valid measure of quality (Neis, Zielstra & Zipf, 2012; Girres & Touya, 2010). However, commonly, such reference data sets are unavailable, as is the case in the Mekong River Delta. As a response, Barron, Neis & Zipf (2014) have attempted to developed a method of intrinsically measuring the quality of OSM data, without the use of a reference set. They do so by for example looking at the historical development of OSM data in a particular region. Since performing such an intrinsic quality assessment for the Mekong River Delta would constitute a dedicated research project in itself, more generic literature was reviewed. Chehreghan and Abbaspour (2018) performed an assessment of the completeness of OSM data. They found that, in comparison to their reference data sets, 87% of objects in the reference set were contained in the OSM database. Although they recognize that this number is higher than those found in comparable studies (Koukoletsos, Haklay, & Ellul, 2012), they note their results are not unexpected, since OSM has undergone an exponential growth of data in recent years (Mooney and Minghini, 2017), and recently developed machine learning methods to improve potential errors in the data sets have proven to be successful (Jilani et al., 2019; Brovelli et al., 2017).

Seeing this results, it was concluded that the use of POI data is indeed suitable and valid for the purpose of this research, and a selection of categories to be used was made. Within the OSM database, POIs are of one or multiple of the following categories: aerialway, aeroway, amenity, barrier, boundary, building, craft, emergency, geological, highway, historic, leisure, military, natural, office, place, power, public transport, railway, shop, sport, telecom, tourism and waterway. Some categories were not included because, in fact, their content does not actually constitute a set of points, but rather a set of linear connections (barrier, geological, highway, waterway). Others were too sparse for the region in question (natural, power, leisure, military, aerialway, aeroway, craft, emergency, railway). Categories as public transport and telecom were not so much sparse overall, but rather sparse in some regions, while containing a vast amount of information in other regions. Due to this

imbalance, these categories were discarded too. Lastly, the categories amenity and place are highly generic. Instead of using these full categories, respective subcategories have been used, allowing a better insight into what features are important to the algorithm. Then, this leads to a set of POI categories used, included in Appendix A.

2.2 Data Preparation

While wealth distribution numbers are published and available for the provinces of Vietnam, it is the aim of this research to produce such numbers on a smaller, local scale. The data described in section 2 will be used as an input for the Random Forest algorithm. In order to produce an output on a local scale, the input data should be local too. This will be achieved by dividing the Mekong River Delta in a raster of local size, each with information relating to the POI and spatial data. After plotting the POIs, a grid with cell size of 5km x 5km has been chosen, resulting in some 2,000 cells. This grid size was chosen since the validation data, detailed in section 2.4, is of similar scale. As such, choosing a comparable local grid scale makes for the most valid accuracy comparison possible.

Then, after producing the grid, the collected data is assigned to the cells in an appropriate manner, described as follows. First, for raster data, represented as quantities, the average value per grid cell is taken as a measure. An exception is the land use layer, consisting of 12 distinct categories: for each grid cell, the most prevalent type of land use has been recorded. Second, the distance from each grid centroid to the closest POI of each category is calculated. As a result, each grid cell has attributes 'distance to closest school', 'distance to closest hospital', etc. Third, networks are treated in a similar manner, where the distance from each grid centroid to the closest instance of each network is calculated. Following, any potentially missing data will be filled in, in an appropriate manner.

As a final step, the dependent variable (poverty level) has to be defined for each grid cell. The only poverty level data available for the Mekong River Delta is on the provincial level. This data is displayed in Appendix B (General Statistics Office of Viet Nam, 2016). Using these average values as a dependent value leads to some heavy assumptions and consequences. First, implicitly, the

expectation is that patterns on provincial level are directly reflected on a local level. Second, as the algorithm to be used aims to predict the values with complete accuracy, a perfectly-performing algorithm will try to return the average value for each province, for all grids in said province. Third, since there is little variation in poverty levels on a provincial level, this will be reflected in the outcomes of the algorithm. The literature analysed does not seem to believe these concerns discredit their methods and obtained results. Rather, it is seen as central to the technique of dasymetric mapping that is used (Stevens et al., 2015; Gaughen et al., 2016; Jean et al., 2016).

2.3 Random forest

In order to predict poverty levels, a Random Forest (RF) algorithm will be used. A RF is an example of an ensemble machine learning method: it combines weaker predictors to produce one, strong predictor (Breiman, 2001). In the case of RF, these weak predictors come in the form of decision trees, which split data into subsets, for each of which a particular prediction should be performed (Swain and Hauska, 1977). A set of randomly grown decision trees is then created, which form the random forest. Advantages of RFs include their fitting speed, the fact that they always converge, and their allowance for automatic parameter tuning, while their lack to predict beyond the range in the training data is usually perceived as a weakness (Wiener and Liauw, 2002). In other words, if training data consists of poverty levels between 5% and 10%, the RF will predict values only within that range, while values such as 4.6% or 10.3% are not completely unlikely, either. However, at the same time, this can also be regarded as a strength of the algorithm: it will never produce values in different orders of magnitude than the input values. As such, though the predictions may be somewhat conservative, they are valid to the degree that they do not show a great deviation from the actual data.

Fitting of the model and estimation of the data were performed using the python-based *sklearn* library, providing a *RandomForestRegressor* package. In this case, a regressor is used (as opposed to a classifier), due to the continuous nature of the dependent variable. The RF is trained on a

subset of the data (the training set), after which it is used to estimate the remainder of the data (the test set). This way, the regressor will only encounter 'new' data, reducing the tendency of overfitting to the encountered data. Two parameters of the forest need to be set: the number of estimators (trees), and the maximum depth of the tree. The function *GridSearchCV*, which iterates over a range of potential values and returns those with the highest R squared score, was employed to find the ideal values for these parameters. As a result, values of 13 and 100 for number of estimators and maximum depth, respectively, were found. Due to the relatively small size of the data set (~2.000 features with ~35 attributes), the standard RF package is able to fit the full set in less than a second.

The trained random forest regressor is then used to predict poverty levels on a 5km x 5km grid scale. As an input, the matrix with attributes per feature is entered. Of all of the trees contained in the trained random forest, a tree will be chosen such that the mean squared error (MSE) is minimized. The algorithm produces an output value for each of the 2.000 cells. These values are used to produce a dasymetric map, displaying poverty level distributed over cells in the Mekong River Delta.

2.4 Accuracy tests

When poverty level data (or any form of data, for that matter) is unavailable on small scale for complete region, a subregion for which such data is available is usually used to validate the results of a prediction (Stevens et al., 2015; Gaughen et al., 2016). The only data available concerning poverty level on a smaller scale, are the poverty levels of 7 districts in the north of the Mekong River Delta (World Bank, 2018). Unfortunately, the district scale is in general still larger than the cell size used in this research. As such, these 7 districts cover some 100 cells in total (<5% of the total amount of cells). Although the data is useful, some side notes must be made. First, due to the scale of the districts, poverty level values are averages for the district, and not for the grid cells. Second, the districts covered in the World Bank report are all from two provinces, Dong Thap and An Giang, located next to each other. Third, all districts are of similar type: rural, with little urban area and located close to the border with Cambodia. As a result, there is no representation of districts on the

low side of the poverty level spectrum: the lowest level for these districts is 5.99%, while the average poverty level for some of the provinces is as low as 1.7%. Ideally, the districts would be a balanced representation of the delta. Since this is not entirely the case, some caution should be taken when interpreting the accuracy found using this data: it is an indication, rather than an exact measure.

3. Results

Using province-level poverty level data from 2016, a map with local (5km x 5km grid) poverty level data has been developed, for the entirety of the Mekong River Delta region. The map colour represents the predicted average poverty level per grid cell, ranging from 1,8% to 10%. As expected, the range of values is exactly equal to the range of values of the dependent variable, is used as input. Figure 1a displays the provincial-level poverty data, while Figure 1b depicts the results of the algorithm.

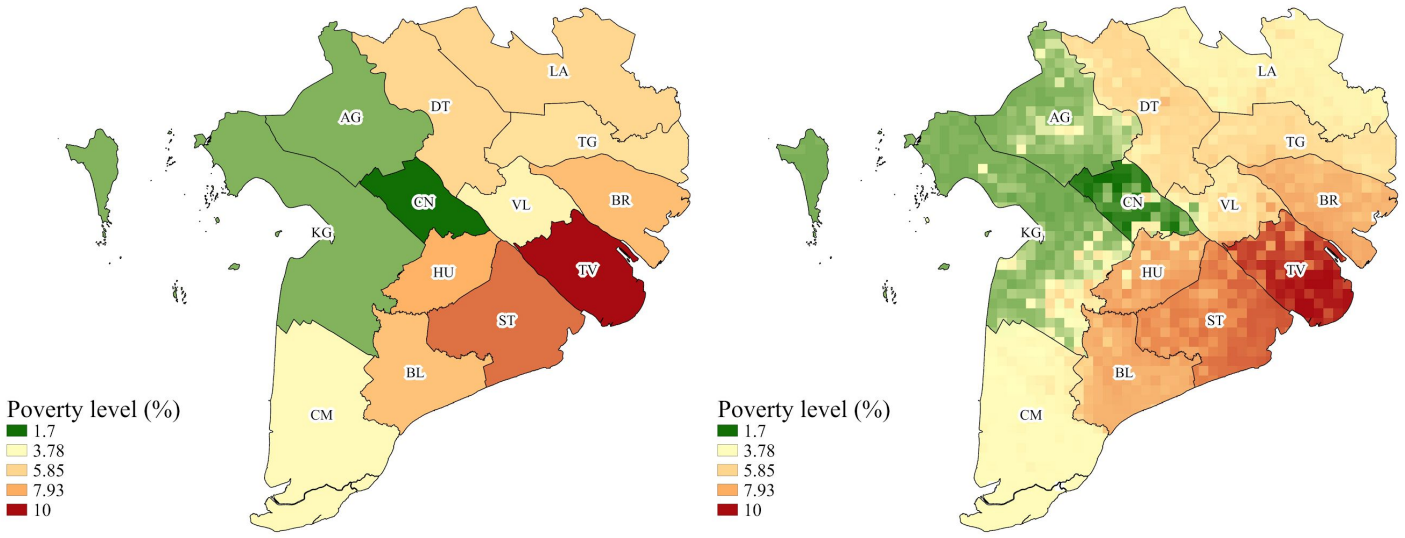

Figure 1a: Map of provincial-level poverty data
Figure 1b: Map of Random Forest prediction for the Mekong River Delta

Although there is some variation in poverty level in selected parts of selected provinces, in general, predicted poverty level values for grid cells are close to the average poverty level of the province they are part of. This is mainly the case for provinces with little POI information, such as Ca Mau (marked *CM*). For example, in this province, predicted poverty levels fluctuate between 3,8% and 4,3%. Here, the algorithm fails to differentiate between different cells: the *distance-to* features all have similar values, when few POIs are located in the province. This is problematic, because there is

an incredibly low probability that all localities within one province have a similar poverty level, as demonstrated by the selection of districts with poverty level data: even though they are located in close proximity, their values are vastly different. At the same time, it must be noted that, where POI data is plentiful, the algorithm appears to be able to differentiate well between cells. Predicted values for Kien Giang (marked *KG*), for example, range between 2,3% and 6,6%; a considerable variation from the mean value of 2,7%.

3.1 Random Forest

The trained random forest reveals information about which features are considered important to the poverty level prediction, and which are not. The twenty most important features are displayed in Figure 2. For each feature, the percentage of the prediction which is explained through this feature, is displayed. The feature importances for all features thus sum up to 100%

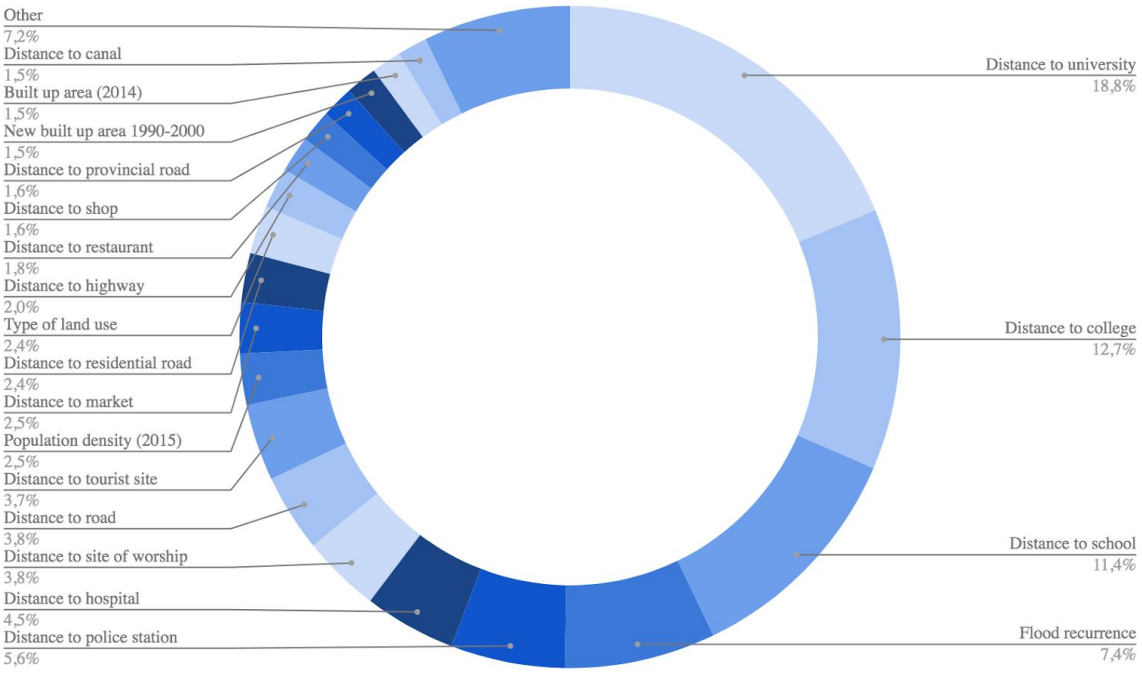

Figure 2: Most important features of RF model

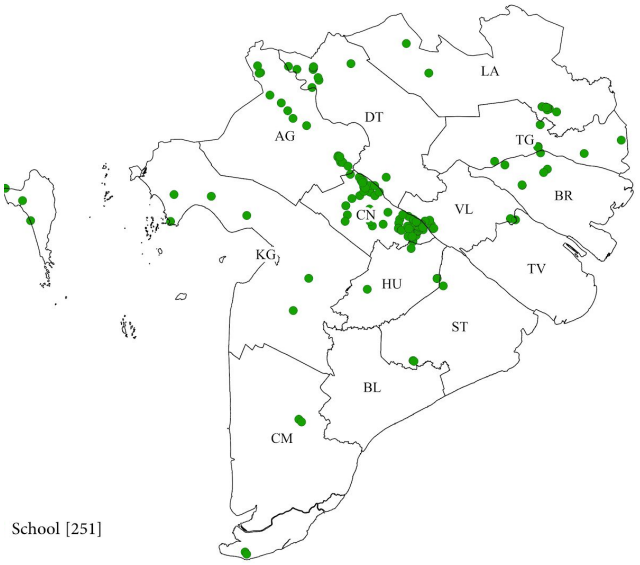

Figure 3: Distribution of school POI

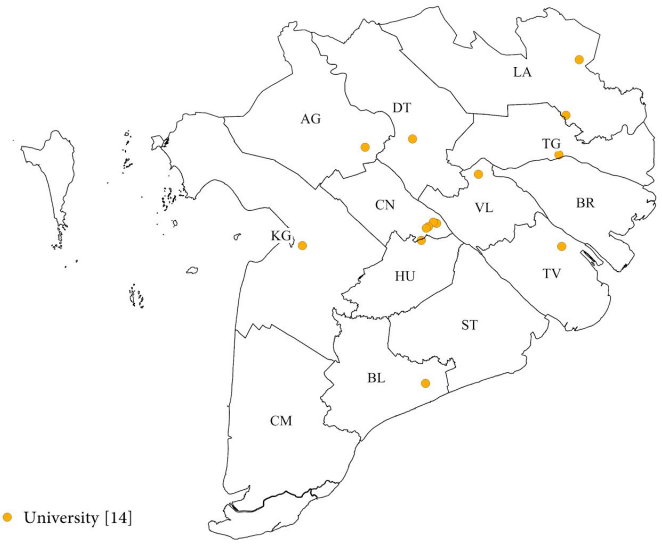

Figure 4: Distribution of university POI

Of the twenty most important features, a majority consists of distance-to-POIs. The distance to university feature has an explanatory value almost twice greater than any other feature, while the other educational POIs also have a high importance. Figures 3 and 4 show the distribution of the POIs *school* and *university*.

Eleven out of thirteen provinces contain at least one university, while every province contains at least one school. Universities are concentrated around Can Tho, with five in the direct vicinity. The fact that distance-to-university is the most important feature, indicates that the distribution of universities comes closest to the distribution of poverty level. As universities are usually located in the capital of each province, the poverty level also appears to be a rural-urban divide. A similar distribution holds for schools, although there are 250 recorded in total, in the Mekong River Delta. Even though no province except Can Tho has more than two universities, the difference in number of schools per province varies greatly. Due to the nature of random forests, it is not possible to display and quantify the relationship with which a particular feature influences the predicted value (Breiman,

2001), i.e. for exameple whether more hospitals closer by lead to a higher poverty level or a lower poverty level.

Other POIs that have a significant influence on poverty level are distance-to-police, distance-to-hospital and distance-to-worship, the latter referring to religious sites. Interestingly, whether or not a cell contains a particular POI, or how many of them it contains, does not influence poverty level, it is the distance to said points that does. The same holds for the feature 'POI?', a binary variable indicating the presence of any POI for a grid cell. Although it appears logical that most POIs have a sphere of influence greater than just their grid cell, it is surprising that the presence of any POI has close to no influence on poverty level. For example, it could be hypothesized that government offices are usually located in more affluent areas, and the presence of a government office in a grid cell as a result would lead to a higher poverty level. Based on the retrieved results, this appears false. Stevens et al. (2015), in their effort to predict poverty levels in Kenya and Cambodia come to the same conclusion: their distance-to features also have a high predictive power, while the presence of a POI in a particular cell does not. Ye et al (2019), however, have contradictory findings. Their variable *POI Density* is only slightly less powerful than *Distance-to-POI*. Here, it must be noted that they employed a single *Distance-to-POI* variable, as opposed to a *Distance-to* variable for each POI categories, as performed in this research. Additionally, through the use of Baidu, their POI coverage is near-complete (Ye et al., 2018). The discrepancy in results may be explained by these two factors.

A second category of variables with high explanatory power are those related to infrastructure networks, specifically the road, highway, provincial road and canal networks, with variables distance_to_road, distance_to_highway, distance_to_provincial_road and distance_to_canal representing the distance from each cell centroid to the nearest element of these networks.

Third, some of the variables related to share of land built-up have some influence on the prediction. Specifically, the variables representing mean built-up share in 2014, and the growth of built-up share between 1990 and 2000. It is hard to interpret why it is exactly these variables that influence the prediction, and not other related to built-up land. Additionally, the mean population

density in 2015 has a limited influence in the algorithm. The predictive power, though limited, of this variable may be explained by the fact that the poverty level data is also from 2015, and there exist some correlation between population density and poverty level.

Lastly, two independent variables with a significant contribution are mean flood recurrence and median land use. The former is informative about the extent to which a piece of land is subject to flooding, over a prolonged period of time. Seeing that flood_mean is the second most important feature, it is suggested that a strong correlation exists between flood recurrence and poverty level. The latter variable represent the most frequent type of land use for each of the grid cells. At present, the algorithm is not informative as to what categories of land use contribute to the prediction of poverty level.

3.2 Accuracy measure

As discussed before, poverty level data for a selection of districts in the Dong Thap and An Giang province will serve as an accuracy measure. The poverty levels for Thanh Bin, Tam Nong, Cao Lanh, An Phu and Tinh Bien are 6,61%, 8,53%, 5,99%, 7,9% and 9,9%, respectively. Table 4 displays the average deviation of the predicted values from the actual value for these districts.

Province	**Dong Thap**			**An Giang**	
Average poverty level	5,8%			2,7%	
District	Thanh Bin	Tam Nong	Cao Lanh	An Phu	Tinh Bien
Average deviation from mean poverty level	1,18%	3,18%	0,51%	4,61%	7,12%
Average deviation as a percentage of mean poverty level	20,3%	54,8%	8,8%	170,7%	263,7%

Table 4: accuracy measures based on selection of districts

Based on this small study, it appears that values predicted for Dong Thap are in general closer

to the actual value, than values predicted for An Giang. This can partially be explained through the dependence of the model on average poverty level values for the provinces. Dong Thap, namely, has an average poverty level of 5,8%, while the same measure is 2,7% for An Giang (General Statistics Office of Vietnam, 2016). The districts for which data is available on a smaller scale all have poverty levels greater than these averages, and the deviation for the districts in An Giang is particularly big. Due to the conservativeness of the RF model, predictions for cells, in general, do not show extreme deviations from the mean of the province. While this is true for most provinces, this is particularly visible in provinces with little POI data, where the algorithm fails to differentiate between the cells. While the deviations found in Dong Thap may be permissible, those found in An Giang (up to 7%) are problematic.

It is possible to assess these results when comparing then to the body of literature similarly employing Random Forests with the aim of dasymetric mapping, be it concerning poverty or population. Stevens et al. (2015) have provided the root mean square error (RMSE) and the root mean square error as a percentage of the mean population size of the cell analyzed (%RMSE) in their population mapping research. The latter measure, %RMSE, gives an indication of the deviation from the actual value. In the case of Vietnam, they conclude a RMSE of 61,92% of the mean population size. Gaughen et al. (2016), while producing spatiotemporal patterns of population in mainland China found an average %RMSE of approximately 95% for each of the different years for which they produced a map, based on a comparison with local data on a small subset of China. In their effort to improve upon this research, Ye et al. (2018) retrieve values of approximately 70% for the same area of focus. In light of these results, it may be concluded that the algorithm produced in this research performs better than the body of literature when focusing on the Dong Thap province, but worse in An Giang. It must be noted that the aforementioned researchers all possessed a larger local data set with which they could perform their accuracy assessment. Thus, while their %RMSE values may thus be assumed to be representative of the entire region under survey, the average deviation values found in this research may not be..

4. Discussion

This research has proposed a method of downsampling provincial poverty data to a local scale, using a Random Forest algorithm and openly and freely available ancillary data. This section will first outline what shortcomings this approach has, focusing on the data used and the interpretation of the results, after which recommendations for future extensions of this research will be described.

4.1 Data

A wide variety of data has been used as the input to the Random Forest algorithm, from a variety of sources. Considerations here are the completeness and accuracy of each data source and potential correlation between data. It has been shown that some data sets, in particular those derived from OSM are not entirely complete. In general, experience from OSM data in the Mekong River Delta shows that the more rural an area is, the fewer points are recorded. To a degree, this may be a reflection of reality. However, even urban centers in otherwise rural provinces that are hard to reach, barely contain any information. These concerns have previously been described in section 2.1. Due to the nature and sources of the other data sets used, it has been assumed that these are complete and valid. An additional point of focus is that correlation may exist between different data sources. It is not unlikely, for example, that a relationship exists between percentage built-up and population density. However, this does not necessarily pose a problem: the RF algorithm is assumed to be insensitive to such correlations, which thus will not impact the predictions (Wiener, 2002). In that sense, there can never be 'enough' data; the algorithm will never predict worse when more data is added.

4.2 Interpretation

One of the challenges of using a Random Forest model for prediction, is the fact that the algorithm tends towards a 'black box': to a degree, it is unclear in what manner the algorithm has come

to its predictions (Breiman, 2001). As a result, it is not always easy to interpret results of the algorithm. As discussed previously, it is possible to get an insight into *what* features are important predictors, but not exactly *how* they are important: due to the nature of RF, it is impossible to describe a direct relationship between the value of the variable, and the predicted value. Then, in order to better understand the complex processes underlying the prediction, it is vital to find a method of interpreting them. A frequently employed example of such a method may be a Gradient Boosting algorithm, for example *GradientBoostingRegressor* by sklearn (Pedregosa et al., 2011). Such algorithms, while similar to other forms of machine learning in the sense that they can form predictions based on a large set of data, store information about the relationship between the value of a independent variable, and the resulting prediction (Shapire, 2003). Problematically, these are relationships within the Gradient Boosting algorithm, that do not necessarily translate equivalently in the Random Forest algorithm. Therefore, these relationships should be taken more as a general trend, than the exact relationship between input and output within the Random Forest algorithm.

4.3 Future directions

There are multiple measures that can be taken to produce a more robust poverty level prediction for the Mekong River Delta, potentially applicable to other regions as well. First, most obviously, and previously discussed, improved and more accurate base data is expected to lead to a better prediction. Next, in order to verify results obtained, it is vital to have a larger set of local census data related to poverty level. This data set should ideally be equally distributed across the Mekong River Delta, and consist of the smallest administrative level. Gaughan et al. (2016), for example, obtained local population density data for four major Chinese metropolitan areas to assess their predicted population density data for the entirety of mainland China. A worthwhile alternative approach to the problem of local poverty mapping may be to use a subset of available local data as a training set for a machine learning algorithm, and using this to produce such local data for areas where no information is available. This constitutes a drastically different approach than dasymetric mapping,

and although it comes with its own limitations and concerns, it may be an alternative that is worth exploring. To our best knowledge, no such research applied to population or poverty mapping has been performed at present.

Moreover, an improvement could be made to the manner in which the networks and POIs are currently treated. Whereas distance to a network or point may be a valuable variable, 'access' can be measured in a better way by variables time-to-network and time-to-POI, recording the time it takes to reach a particular network or destination. When someone lives in short distance of a highway, but their path there is obstructed by a waterway, this network may not be of use to them at all. Such variables can be incorporated, again with the help of OpenStreetMap (Huber and Rust, 2016). The command developed by Huber and Rust (2016), called *osrmtime*, allows one to compute travel time by foot, car or bike between any two points around the world. A point of caution here is the computational power required to compute the travel time between any two points in a region. As the data used in these tools solely consists of the road network and governmentally imposed speed limits, they may not be an accurate reflection of actual travel time: deviations may occur due to congestion or poor road quality. As an extension, crowd-sourced trajectory mining may be employed (Basari et al., 2016). This technique relies on data from for example speeding cameras, registering the average speed on certain sections of road. Although Basari et al. (2016) note that there is a high chance of anomalies in such data, they pronounce confidence in existing techniques to detect such anomalies. Therefore, they conclude that crowd-sourced trajectory mining can be used to form a more robust data set of travel times.

Additionally, variables used at present can be broken down into more subcategories. Data regarding land use, for example, has been recorded as the most frequently occurring type of use for each cell. While this measure gives a valuable indication of land use, one also loses some of the meaning while using this form of upscaling. An alternative approach would be, to create a binary variable for each type of land use. In other words, for each grid cell, we would record whether or not it

includes a particular form of land use. Alternatively one could record the proportions of land use within each grid cell.

Lastly, the case study performed on the Mekong River Delta using a Random Forest algorithm may be extended through the use of a Convolutional Neural Network (CNN) machine learning algorithm. A limited set of literature has employed this method, and results are generally promising (Jean et al., 2016; Perez et al., 2017). As discussed previously, however, CNNs constitute a 'black box' even more so than a Random Forest algorithm. Since the RF algorithm has succeeded in providing some insight into important factors relating to poverty distribution of the Mekong River Delta, a suitable next step would be the use of a CNN algorithm. Although through the execution of such algorithm we lose knowledge about the manner in which predictions are made, it has the promise of leading to more accurate results (LeCun & Kavukcuoglu, 2010).

5. Conclusion

This research has proposed and implemented an algorithm to produce local poverty data for the Mekong River Delta, using POIs and ancillary data. Although the implementation was successful, some reservations should be held with regards to its validity. Since validation data is relatively sparse, it is difficult to provide a hard measure of its performance and quality. Additionally, some doubt exists surrounding the completeness of some of the data sets used, specifically the POI data. The suggestion has been made that in future research, when high-quality local poverty data is available for a subset of an area, a different approach may be to train the algorithm on exactly this data, as opposed to higher-level poverty data. Based on this case study in the Mekong River Delta, it follows to conclude that in general, when certain conditions and assumptions are met, employing a Random Forest algorithm appears a worthwhile form of producing local poverty level data. Specifically, it is assumed that wealth is distributed locally in a manner similar to on a provincial level, and the data used should be complete and of high quality. Further research in other areas, where such data is available, will be able to determine whether the results found in this paper can be extrapolated to other regions, and ultimately be accepted as a general truth, irrespective of locality.

References

Bakillah, M., Liang, S., Mobasheri, A., Jokar Arsanjani, J., & Zipf, A. (2014). Fine-resolution population mapping using OpenStreetMap points-of-interest. *International Journal of Geographical Information Science*, *28*(9), 1940–1963. https://doi.org/10.1080/13658816.2014.909045

Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic OpenStreetMap quality analysis. *Transactions in GIS*, *18*(6), 877-895.

Basiri, A., Jackson, M., Amirian, P., Pourabdollah, A., Sester, M., Winstanley, A., ... & Zhang, L. (2016). Quality assessment of OpenStreetMap data using trajectory mining. *Geo-spatial information science*, *19*(1), 56-68.

Bennett, M. M., & Smith, L. C. (2017). Advances in using multitemporal night-time lights satellite imagery to detect, estimate, and monitor socioeconomic dynamics. *Remote Sensing of Environment*, *192*, 176–197. https://doi.org/10.1016/j.rse.2017.01.005

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Brovelli, M. A., Minghini, M., Molinari, M., & Mooney, P. (2017). Towards an automated comparison of OpenStreetMap with authoritative road datasets. *Transactions in GIS*, *21*(2), 191-206.

Channan, S., Collins, K., & Emanuel, W. R. (2014). Global mosaics of the standard MODIS land cover type data. *University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA*, *30*.

Chehreghan, A., & Ali Abbaspour, R. (2018). An evaluation of data completeness of VGI through geometric similarity assessment. *International Journal of Image and Data Fusion*, *9*(4), 319-337.

Elvidge, C. D., Sutton, P. C., Ghosh, T., Tuttle, B. T., Baugh, K. E., Bhaduri, B., & Bright, E. (2009). A global poverty map derived from satellite data. *Computers & Geosciences*, *35*, 1652–1660. https://doi.org/10.1016/j.cageo.2009.01.009

Gaughan, A. E., Stevens, F. R., Huang, Z., Nieves, J. J., Sorichetta, A., Lai, S., … Tatem, A. J. (2016). Spatiotemporal patterns of population in mainland China, 1990 to 2010. *Scientific Data*, *3*, 1–11. https://doi.org/10.1038/sdata.2016.5

Girres, J. F., & Touya, G. (2010). Quality assessment of the French OpenStreetMap dataset. *Transactions in GIS*, *14*(4), 435-459.

Hoang, L. P., Biesbroek, R., Tri, V. P. D., Kummu, M., Vliet, M. T. H. Van, Leemans, R., … Ludwig, F. (2018). Managing flood risks in the Mekong Delta : How to address emerging challenges under climate change and socioeconomic developments. *Ambio*, *47*(6), 635–649. https://doi.org/10.1007/s13280-017-1009-4

Huber, S., & Rust, C. (2016). Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM). *The Stata Journal*, *16*(2), 416-423.

Jean, N., Burke, M., Xie, M., Matthew Davis, W., B. Lobell, D., & Ermon, S. (2016). Combining satellite imagery and machine learning to predict poverty. *Science*, *353*(6301), 790–794.

Koukoletsos, T., Haklay, M., & Ellul, C. (2012). Assessing data completeness of VGI through an automated matching procedure for linear data. *Transactions in GIS*, *16*(4), 477-498.

Jilani, M., Bertolotto, M., Corcoran, P., & Alghanim, A. (2019). Traditional vs. Machine-Learning Techniques for OSM Quality Assessment. In *Geospatial Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 469-487). IGI Global.

LeCun, Y., & Kavukcuoglu, K. (2010). Convolutional Networks and Applications in Vision. In *Proceedings of 2010 IEEE International Symposium on Circuits and Systems* (pp. 253–256). Paris. https://doi.org/10.1109/ISCAS.2010.5537907

Ma, T., Zhou, Y., Zhou, C., Haynie, S., Pei, T., & Xu, T. (2015). Night-time light derived estimation of spatio-temporal characteristics of urbanization dynamics using DMSP / OLS satellite data. *Remote Sensing of Environment*, *158*, 453–464. https://doi.org/10.1016/j.rse.2014.11.022

Mooney, P., & Minghini, M. (2017). A review of OpenStreetMap data. 37–59. https://doi.org/10.5334/bbf.c

Mellander, C., Lobo, J., Stolarick, K., & Matheson, Z. (2015). Night-Time Light Data : A Good Proxy Measure for Economic Activity ? *PLoS ONE*, *10*(10), 1–18. https://doi.org/10.1371/journal.pone.0139779

Neis, P., Zielstra, D., & Zipf, A. (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, *4*(1), 1-21.

Njuguna, C., & McSharry, P. (2017). Constructing spatiotemporal poverty indices from big data ☆. *Journal of Business Research*, *70*, 318–327. https://doi.org/10.1016/j.jbusres.2016.08.005

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. *Journal of machine learning research*, *12*(Oct), 2825-2830.

Perez, A., Yeh, C., Azzari, G., Burke, M., Lobell, D., & Ermon, S. (2017). Poverty Prediction with Public Landsat 7 Satellite Imagery and Machine Learning. In *31st Conference on Neural Information Processing Systems* (pp. 1–6). Retrieved from http://arxiv.org/abs/1711.03654

Robnik-Šikonja, M. (2004). Improving Random Forests. *Lect. Notes Comput. Sc.*, (3201), 359–370. https://doi.org/10.1007/978-3-540-30115-8_34

Stevens, F. R., Gaughan, A. E., Linard, C., & Tatem, A. J. (2015). Disaggregating census data for population mapping using Random forests with remotely-sensed and ancillary data. *PLoS ONE*, *10*(2), 1–22. https://doi.org/10.1371/journal.pone.0107042

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear estimation and classification*(pp. 149-171). Springer, New York, NY.

Swain, P. H., & Hauska, A. N. D. H. (1977). The Decision Tree Classifier : Design and Potential. *IEEE Transactions on Geoscience Electronics*, *15*, 142–147. https://doi.org/10.1109/TGE.1977.6498972

Wiener, Matthew; Liaw, A. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. https://doi.org/10.1177/154405910408300516

Winsemius, H. C., Jongman, B., Veldkamp, T. I. E., Hallegatte, S., Bangalore, M., & Ward, P. J. (2018). Disaster risk, climate change, and poverty: assessing the global exposure of poor people to floods and droughts. *Environment and Development Economics*, *23*, 328–348. https://doi.org/10.1017/s1355770x17000444

World Bank. (2018). *Climbing the Ladder: Poverty Reduction and Shared Prosperity in Vietnam*.

Ye, T., Zhao, N., Yang, X., Ouyang, Z., Liu, X., Chen, Q., … Jia, P. (2019). Improved population mapping for China using remotely sensed and points-of-interest data within a random forests model. *Science of the Total Environment*, *658*, 936–946. https://doi.org/10.1016/j.scitotenv.2018.12.276

Appendix A: Overview of OSM POIs used

Name	OSM Query	Count	Notes
College	amenity=college	28	
Hospital	amenity=hospital	88	no entries for some provinces
Marketplace	amenity=marketplace	62	good spatial distribution
Office	office=*	278	public offices, e.g. town hall
Police	amenity=police	43	
Restaurant	amenity=restaurant	558	
Religious site	amenity=place_of_worship	133	
Shop	shop=*	1.283	includes all shop forms registered in OSM
School	amenity=school	251	primary and secondary schools, no entries for some provinces
Tourism	tourism=*	1.019	
University	amenity=university	4	all in Can Tho
Total		3.747	

Appendix B: Average poverty levels per province

Province	**Average poverty level**
Dong Thap	5,8%
Long An	4,2%
Tien Giang	5,3%
Ben Tre	7,1%
Tra Vinh	10%
Vinh Long	4,3%
An Giang	2,7%
Kien Giang	2,7%
Can Tho	1,7%
Hau Giang	7,7%
Soc Trang	8,7%
Bac Lieu	6,9%
Ca Mau	4,0%

Average poverty levels per provinces of the Mekong River Delta (General Statistics Office of Viet Nam, 2016)

Appendix C: Python code used

```
import pandas as pd
import numpy as np

# Import data
data = pd.read_csv("test9.csv")

# Display first five elements of the data
data.head()

# We make a copy of the data, leaving the original dataFrame
intact
datas = data.copy()

# We set the column 'grid_id' as the index of the dataFrame
datas = datas.set_index("grid_id")

# Create an empty 'holding' column for y-values
datas['y'] = np.nan

# We shuffle the data
datas = datas.sample(frac=1)

# Insert average poverty rate per province for grid cells in these
# respective provinces

for index, row in datas.iterrows():
    if row["PROV"] == 'DT':
        datas.at[index, 'y'] = 0.058
    if row["PROV"] == 'LA':
        datas.at[index, 'y'] = 0.042
    if row["PROV"] == 'TG':
        datas.at[index, 'y'] = 0.053
    if row["PROV"] == 'BR':
        datas.at[index, 'y'] = 0.071
    if row["PROV"] == 'TV':
        datas.at[index, 'y'] = 0.100
    if row["PROV"] == 'VL':
        datas.at[index, 'y'] = 0.043
    if row["PROV"] == 'AG':
        datas.at[index, 'y'] = 0.027
    if row["PROV"] == 'KG':
        datas.at[index, 'y'] = 0.027
    if row["PROV"] == 'CN':
        datas.at[index, 'y'] = 0.017
    if row["PROV"] == 'HU':
        datas.at[index, 'y'] = 0.077
    if row["PROV"] == 'ST':
        datas.at[index, 'y'] = 0.087
    if row["PROV"] == 'BL':
```

```
        datas.at[index, 'y'] = 0.069
    if row["PROV"] == 'CM':
        datas.at[index, 'y'] = 0.040

# Define the X matrix, consisting of a selecting of columns from
the
# dataFrame
X = datas[['flood_mean', 'total_pts', 'LandUse_ma', 'Canal_dist',
           'Highway_di', 'Prov_dist', 'Resi_dist', 'Road_dist',
           'New_road_d', 'CollegeDis', 'HospDist', 'MarketDist',
           'OfficeDist', 'PoliceDist', 'RestDist', 'SchoolDist',
           'ShopDist', 'TourDist', 'WorshDist', '15pop_mean',
           '10pop_mean', '10_15_diff', '90pop_mean', '75pop_mean',
           '75_90_diff', '90_10_diff', '14_built_m', '75_built_m',
           '90_built_m', '00_built_m', '75_90_blt', '90_00_blt',
           '00_14_blt', 'POI?', 'Uni2Dist']].copy()

# The y array is the last column of the datas dataFrame
y = datas['y'].copy()
y.head()

# Divide train and test set
from sklearn.model_selection import train_test_split
X_train,   X_test,   y_train,   y_test   =   train_test_split(X,   y,
test_size=0.3, random_state=42)

train = datas[datas['y'].notnull()]
test = datas[datas['y'].isnull()]

X_train = train[["flood_mean", "CollegeDis", "HospDist",
                 "MarketDist", "OfficeDist", "PoliceDist",
                 "RestDist", "SchoolDist", "ShopDist","TourDist",
                 "WorshDist", "total_pts", "LandUse_ma",
                 "Canal_dist", "Highway_di", "Prov_dist",
                 "Road_dist", "New_road_d", "15pop_mean",
                 "10pop_mean", "10_15_diff", "75pop_mean",
                 "90pop_mean", "75_90_diff", "90_10_diff",
                 "75_built_m", "90_built_m", "00_built_m",
                 "14_built_m",       "75_90_blt",       "90_00_blt",
"00_14_blt",
                 "Uni2Dist"]].copy()

y_train = train['y'].copy()

X_test = test[["flood_mean", "CollegeDis", "HospDist",
                 "MarketDist", "OfficeDist", "PoliceDist",
                 "RestDist", "SchoolDist", "ShopDist","TourDist",
                 "WorshDist", "total_pts", "LandUse_ma",
                 "Canal_dist", "Highway_di", "Prov_dist",
                 "Road_dist", "New_road_d", "15pop_mean",
                 "10pop_mean", "10_15_diff", "75pop_mean",
```

```
                "90pop_mean", "75_90_diff", "90_10_diff",
                "75_built_m", "90_built_m", "00_built_m",
                "14_built_m",       "75_90_blt",       "90_00_blt",
"00_14_blt",
                "Uni2Dist"]].copy()

y_test = test['y'].copy()

# Fit regressor
from sklearn.ensemble import RandomForestRegressor
regr    =    RandomForestRegressor(max_depth=13,    random_state=0,
n_estimators=100)
regr.fit(X_train, y_train)

pred = regr.predict(X)

pred

X['pred'] = pred
pd.DataFrame(X[['pred']]).to_csv("pred.csv")

# Find R squared
from sklearn.metrics import r2_score
r2_score(y_test, pred)

# Display most important features
feat_importances = pd.Series(regr.feature_importances_,
                             index=X.columns)
feat_importances.nlargest(20).plot(kind='barh'
)
```