Public Disclosure Authorized

Public Disclosure Authorized

Policy Research Working Paper 9979

Where Are All the Jobs?

A Machine Learning Approach for High Resolution Urban Employment Prediction in Developing Countries

Samira Barzin
Paolo Avner
Jun Rentschler
Neave O'Clery

WORLD BANK GROUP
Urban, Disaster Risk Management, Resilience and Land Global Practice
March 2022

Policy Research Working Paper 9979

Abstract

Globally, both people and economic activity are increasingly concentrated in urban areas. Yet, for the vast majority of developing country cities, little is known about the granular spatial organization of such activity despite its key importance to policy and urban planning. This paper adapts a machine learning based algorithm to predict the spatial distribution of employment using input data from open access sources such as Open Street Map and Google Earth Engine. The algorithm is trained on 14 test cities, ranging from Buenos Aires in Argentina to Dakar in Senegal. A spatial adaptation of the random forest algorithm is used to predict within-city cells in the 14 test cities with extremely high accuracy (R- squared greater than 95 percent), and cells in out-of-sample "unseen" cities with high accuracy (mean R-squared of 63 percent). This approach uses open data to produce high resolution estimates of the distribution of urban employment for cities where such information does not exist, making evidence-based planning more accessible than ever before.

This paper is a product of the Urban, Disaster Risk Management, Resilience and Land Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at samira.barzin@maths.ox.ac.uk, pavner@worldbank.org, jrentschler@worldbank.org, and oclery@maths.ox.ac.uk.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Produced by the Research Support Team

Where Are All the Jobs? A Machine Learning Approach for High Resolution Urban Employment Prediction in Developing Countries

Samira Barzin[a,*], Paolo Avner[c], Jun Rentschler[c] and Neave O'Clery[a,b]

a Mathematical Institute, University of Oxford
b The Bartlett Centre for Advanced Spatial Analysis, University College London
c The World Bank

Corresponding Author: *samira.barzin@maths.ox.ac.uk,

Keywords: development economics, urban economics, cities, firm locations, big data, satellite data, computational methods, machine learning
JEL Classification: O18, C6, R3, R11

Acknowledgements: The authors are grateful to Tatiana Peralta Quirós, who was involved in an early effort to develop a methodology to predict employment distribution in urban areas and was of great help in accessing employment datasets. We also thank Holly Krambeck, Nancy Lozano, Lorenzo Carrera, Samuel Heroy and Sam Fankhauser for helpful feedback on an earlier version. This study was supported by the Global Facility for Disaster Reduction and Recovery (GFDRR) and by the Foreign, Commonwealth and Development Office (FCDO) through the Multi-donor Trust Fund on Sustainable Urbanization, and the Oxford Martin Programme on Informal Cities. The computational aspects of this research were supported by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z and the NIHR Oxford BRC. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

1 Introduction

Cities across developing countries have experienced unparalleled population growth in recent decades, fueled both by rural-urban migration and high population growth (Henderson and Kriticos (2018), Bryan et al. (2020)). These dynamics differ substantially from traditional urban growth patterns experienced by European and North American cities, particularly in terms of speed and the noticeable absence of structural transformation and industrialization (Henderson and Kriticos, 2018). In general, urbanization is strongly associated with economic growth on a national scale (Glaeser and Xiong, 2017) and agglomeration economies (economic benefits arising from spatial clustering of economic activity) (Bryan et al., 2020). However, negative externalities arising from high urban density, such as congestion and insufficient sanitation infrastructure, require well planned policy action in terms of, for example, infrastructure and institutions. It is thus crucial to improve our understanding of the dynamics and inner functionalities of cities in order to design more targeted policies to support successful urban development. In particular, gaining an insight into how jobs are spatially distributed is key to understanding urban form and internal labor market structure.

These assessments matter. Metrics that estimate accessibility to jobs aim to capture in a simple fashion how an urban area's transport systems and land use perform in connecting workers to employment opportunities, a core function of cities (Duranton and Puga, 2015). It is also useful to uncover areas of a city that are underserved by public transport systems, where residents have low access to jobs (i.e., spatial mismatch) (Gobillon and Selod, 2014) and where public interventions would be beneficial. In fact, employment accessibility studies have become a mandatory pre-requisite for World Bank financed urban transport operations in recent years. In parallel, there is increasing recognition that natural disasters and climate change impose significant costs on societies not only through damage to infrastructure assets but also through disruptions to the infrastructure services that these systems enable (Hallegatte et al., 2019). Disruptions to commuting movements entail large opportunity costs in the form of lost time, discomfort or foregone salaries (He et al., 2021). Minimizing or even preventing these disruptions is possible through targeted maintenance of, or investments in, road infrastructure identified through network criticality analyses, but these hinge on an understanding of job locations. Finally, there is much to gain in terms of furthering our understanding of the urban agglomeration forces at play, particularly with respect to employment density (Ciccone and Hall, 1996) and spatial clustering, i.e. monocentric, polycentric, or dispersed (Lall et al., 2021a; Bertaud, 2002).

However, in most developing country cities, granular information on jobs is often not available as geo-referenced business registries and employment censuses rarely exist or are incomplete and outdated. Given these exercises are costly and burdensome to organize, requiring the mobilization, training and deployment of surveyors, they are only conducted

infrequently and if undertaken often suffer from quality issues. Here we aim to fill this key knowledge gap by developing a new machine learning technique to predict the spatial distribution of employment from openly and widely available data on amenities, roads, light intensity and other variables such as population density.

A burgeoning literature has recently emerged that harnesses satellite data and so-called 'big data' to uncover socio-economic and spatial patterns of cities. Of particular note is a milestone study by Henderson et al. (2012) where the authors show that night light intensity derived from satellite data can be used as a proxy for income/GDP. A second set of studies focus on estimating the spatial extent of urban areas in developing countries. For instance, Henderson et al. (2021b) rely on satellite based population data to identify urban areas and estimate wages for cities in six SSA countries. Focusing on building characteristics, Lall et al. (2021b) and Henderson et al. (2021a) rely on building height data to investigate urban form for 397 cities and identify slum areas in Nairobi respectively. Another valuable source of data is derived via digital (often crowd-sourced) maps, e.g., Open Street Map (OSM). For example, Baruah et al. (2021) combine satellite based population and night lights data with OSM data on the road network to compare the urban form of cities formally under anglophone and francophone colonial rule, while Soman et al. (2020) rely on OSM data to detect informal settlements.

A related strand of literature combines non-traditional data with machine learning approaches in order to predict a range of socio-economic parameters thus overcoming data sparsity challenges in developing countries. For example, Jean et al. (2016) rely on high resolution satellite imagery in combination with convolutional neural network (CNN) analysis to predict wealth across five SSA countries, where CNN is used to extract roads and other features from the raw pixels of the satellite imagery. Yeh et al. (2020) extend this work and predict asset wealth at high resolution for 20,000 villages located across SSA. Similarly, Engstrom et al. (2021) exploit deep learning methods to predict poverty and consumption at the local level across Sri Lanka. Blumenstock et al. (2015) alternatively rely on mobile phone data to infer socio-economic status at an individual level via a machine learning classification analysis. While these studies represent just a fraction of efforts in the literature, they capture an emerging trend in the deployment of novel data and machine learning to fill gaps in data collection in low resource environments.

In terms of novel approaches to specifically estimating employment, the literature is more limited. The 'gold standard' is census data on firms or employees, typically assembled by government agencies in the course of collecting taxes. It is hence normally limited to the formal sector. Travel surveys, in contrast, typically have much more limited coverage of the population, but can often be statistically upweighted to provide a reasonably accurate picture of employment distribution - and may also include the informal sector. However, as is the case for business registries and employment censuses, travel surveys are also scarce

and conducted inconsistently. To overcome this data gap, some scholars have inferred the presence of firms from imagery, including both street imagery (Straulino et al., 2021) and satellites (Goldblatt et al., 2020). However, these approaches are severely limited to picking up just what is 'seen' from the exterior or shape of buildings. Other authors have successfully inferred work locations (among other information) from mobile phone data, either through Call Detail Records, CDRs, (Kreindler and Miyauchi, 2021; Zagatti et al., 2018; Louail et al., 2015) or more granular information obtained from anonymized GPS logs collected from smartphone users through third party apps (Miyauchi et al., 2021; Yabe et al., 2020, 2021). These approaches are promising but have several drawbacks. First, they are time consuming, as obtaining agreement from mobile phone subscription providers or third parties for data use, accessing the data in a secure manner and processing it usually takes several months. Second, accessing this data can be costly when not made freely available for humanitarian purposes (Lu et al., 2012; Bengtsson et al., 2015). Thirdly, there are concerns that this type of data may lack proportional coverage over space, notably for CDRs where cell phone towers are geographically unequally distributed. Similarly, coverage is likely to vary across sub-populations, especially when the data comes in the form of GPS logs obtained from smartphones that have a low penetration share in developing country economies and are mainly owned by higher income groups. Quick to deploy, cheap and scalable alternative methodologies to understand the spatial distribution of jobs in urban areas are urgently needed.

Here, we develop a novel method to predict the spatial density of employment using open freely available data from web-based sources including OpenStreetMap (OSM) and Google Earth Engine (GEE). OSM contains information on the spatial location of hundreds of amenities and other mapped attributes, ranging from bus stops to shops to the street network. These variables hold invaluable information on human activity, which we hypothesize is highly correlated with employment density. We supplement the OSM variables with satellite data, such as population, night lights and land cover, accessed via GEE. In order to predict employment density from OSM and GEE data sources, we deploy a range of machine learning algorithms, focusing in particular on two ensemble tree methods - Random Forest (RF) and Gradient Boosting Machines (GBM) - a Regularized Generalized Linear Model and a linear regression. We select the RF algorithms specifically due to their ability to handle noise in the training data, stemming from a mix of measurement error and incompleteness in both the employment and feature data. We propose a simple extension to these models, and extend our feature space to include the spatial version of each feature based on feature values in neighboring cells, thus accounting for local spatial clustering. We additionally include Regularized Generalized Linear Models and linear regression based approaches, which while less suited to our noisy data, enable us to delve deeper into variable importance scores, and uncover which features are most indicative of the presence of employment.

Our work differs from previous work along a number of dimensions. First, we focus on

prediction of employment rather than income or urban extent. Second, we predict employment at high spatial resolution within cities, while much similar work focuses on larger spatial scales. Third, we explicitly aim to develop a framework that can predict employment in an out-of-sample setting, where training data for the city in question is not available. This is a key feature, and rarely tackled in the literature, for usability in data scarce environments. Finally, our approach is scalable and more robust than alternative methods, particularly in comparison to deep learning approaches which tend to be less robust to noise in the data. We train the model on data derived from open sources, and use methods that require reasonable and accessible levels of computing power. Our work relates most closely to Goldblatt et al. (2020) who explore the correlation between satellite derived features and employment/enterprise at the local level for Vietnam, and Ahlfeldt et al. (2020) who infer 'prime locations' (defined as locations exhibiting very high density of human capital intensive services) on the basis of Google Places data.

We apply our method to 14 cities, ranging from Buenos Aires in Argentina to Dakar in Senegal. The set of cities was chosen for a variety of reasons including both geographical spread and the availability of granular employment data (for training purposes). We note that the source (and quality) of this employment data varies, and is based on both census and travel surveys. We show that we can predict held back within-city cells in our test cities with extremely high accuracy (> 95% R^2), and cells in out-of-sample cities with high accuracy (mean 63% with max 80% R^2). The variation in out-of-sample prediction can likely be attributed to a combination of city-specific relationships between employment and our features, and varying degrees of data quality for both the feature and employment data across cities. We investigate this latter effect, finding support in the data for this hypothesis.

We hope our work will be widely deployed for policy and humanitarian purposes. Recently, it has been noted that academic advances in the application of data science methods to development questions has yielded little real world impact to date (Burke et al., 2021). This was attributed in part to difficulties in interpretability. By using minimally processed input data sources, we avoid the complex processing required for extracting objects, for example, from raw satellite imagery data. We hope that this simplification, combined with the open nature of the input data and the widely-used random forest algorithm deployed, reduces interpretation issues and barriers to use.

2 Data

We build the analyses and train the algorithms on a sample set of 14 cities for which we have access to urban employment data with spatial identifiers. This data set consists of 9 cities across Sub-Saharan Africa (SSA) and 5 cities in Latin America (LAT); see Figures 1 (a) and (c). Across cities, the data differs by data collection type and by degree of spatial

Table 1: Overview of Data across Cities

City	Data Type	Year Collected	No. of Polygons	Mean Size of Polygons (km^2)	No. of 500m x 500m Grid Cells
SSA					
Abidjan/CIV	Population Census	2015	307	14.79	21,121
Dakar/SEN	Travel Survey	2015	213	2.51	2,663
Dar es Salaam/TZA			101	17.73	8,436
Douala/CMR	Travel Survey	2018	194	1.37	1,399
Harare/ZWE	Population Census	2012	75	12.50	4,558
Kampala/UGA	Firm Census	2011	184,335*	5.05*	4,721
Kigali/RWA	Firm Census	2011	1,162	0.63	3,566
Kinshasa/COD	Travel Survey	2018	395	3.68	7,021
Nairobi/KEN	Travel Survey	2013	106	6.77	3,532
LAT					
Belo Horizonte/BRA	Firm Census	2019	1,048	3.05	13,687
Bogotá/COL			19,915	0.02	2,190
Buenos Aires/ARG	Firm Census	2004/5	11,822	0.30	16,551
Lima/PER	Population Census	2010	91,875	0.01	4,988
Mexico City/MEX	Firm Census	2019	476,853*	0.33*	4,161

Note: Firm Census refers to firm censuses, establishment censuses or other firm registries, or projections based on these. Data excludes units with missing employment information. * Data was provided in latitude/longitude point locations; data for Kampala was spatially matched to the 2012 Kampala travel survey polygons, data for Mexico City was spatially matched to AGEB polygons

disaggregation; city specific details are provided in Table 1. The data is usually collected either via travel/commuting surveys, extrapolated via population censuses or identified from business registries, where the degree of spatial disaggregation is generally higher for the latter, i.e., resulting in smaller polygon units. For cities where we have access to multiple rounds of data, we rely on the most recently collected and/or most spatially disaggregated version of the data[1].

First, we calculate the spatial employment density for each polygon provided; see Figure 1(b). In order to achieve homogeneity of spatial units across the cities and higher spatial disaggregation overall, we overlay each city with a grid net of 500m x 500m hexagonal cells.

[1] See Appendix A for more information on the data sources.

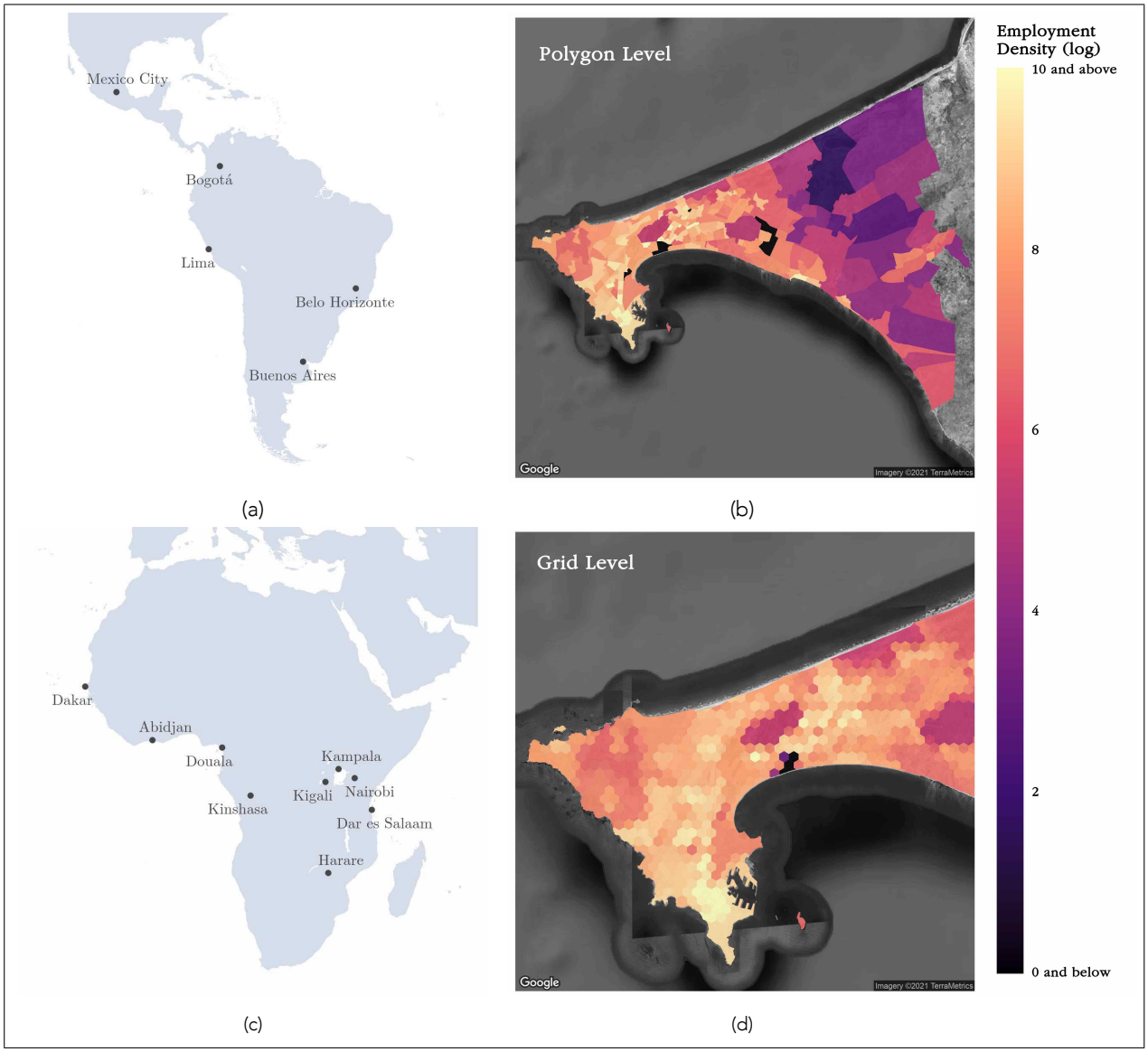

Figure 1: Overview of City Sample and Spatial Distribution of Employment.
Subfigures (a) and (c) depict cities included in the data, while subfigures (b) and (d) illustrate the spatial distribution of employment for original polygons and disaggregated into 500m x 500m hexagonal grid cells for Dakar/Senegal.

We intersect the original polygon and grid cell layers and obtain spatial employment data for each grid cell through weighting employment values by the spatial intersection of grid cells and polygons. For most cities the data is provided as number of jobs per polygon, making the identification of exact job locations impossible. Hence, we assume homogeneity in spatial distribution of employment within polygons when computing grid cell values.

For our 14 cities, we extract OSM data on the cities' major street networks, public transport stations, shop and office locations, amenities, airports, lakes, rivers, parks, forests and coastlines. We further compile satellite data accessed via Google Earth Engine (GEE) on

night lights, population, air pollution, geophysical land surface, built-up land cover, the Normalized Difference Vegetation Index/NDVI, the Normalized Difference Water Index/NDWI and land use[2]. The data on night lights, population, air pollution, NDVI and NDWI contains temporal variation and is thus averaged within its original resolution raster cells across 2015 to 2019 prior to intersecting it with the employment grid cell layer. Both OSM and GEE data is available globally, where OSM is crowd-sourced mapped feature data and thus is assumed to exhibit heterogeneity in data quality and completeness across cities; data included via GEE is assumed to be consistent across included locations. GEE hosts an array of satellite and rasterized data from various sources. For computational arguments and ease of scale-up of the analyses, we limit our choice of satellite data to those available in rasterised form. Consequentially, in contrast to the related literature (see for example Jean et al. (2016) and Yeh et al. (2020)), we do not require deep learning methods or convolutional neural networks to detect and extract features[3] from raw satellite pixels.

For each city, we compute a bounding box around the overall geographical area and extract OSM data for various features located within this box. In order to obtain the grid level values from the mapped features, we count all point features, e.g. shop locations and public transport hubs in a given grid cell. For line features, e.g. roads and rivers, we extract the length per grid cell, and for polygon features, e.g. parks and airports, we calculate the spatial intersection with grid cells. For GEE derived data, feature values per grid cell are computed via a weighted spatial intersection between the original data raster cells and 500m x 500m hexagonal grid cells. Given that cells on the border of the city are cut by the external outline of the city and are thus smaller than 500m x 500m grid cells, we compute and use the density versions of the values for each grid cell for the analyses. In addition to the extracted features, we also derive further features from the raw OSM and GEE data (e.g. street intersections are derived from the OSM data on the city's street network and terrain roughness is computed based on the earth's surface elevation data). Figure 2 provides visualizations of a selection of OSM and GEE features for Dakar/Senegal.

A visual comparison between the images contained in Figure 1 and Figure 2 suggests that both employment and features follow a broadly similar spatial distribution pattern across Dakar. The highest values of both employment and the majority of the features can be observed in the west and in particular the south west of the city. Commercial, administrative and institutional activities in Dakar are concentrated in the "Plateau" and "Fann" neighborhoods, towards the South and South-West of the peninsula. Industrial zones are distributed in a slightly more homogeneous manner between Dakar, Pikine and Rufisque (with the addition of Guédiawaye, the four municipalities, together, constitute the Dakar metropolitan

[2]See Appendix A for further details.

[3]Throughout this work, we use the terms features and covariates, and the terms of response and dependent variable interchangeably.

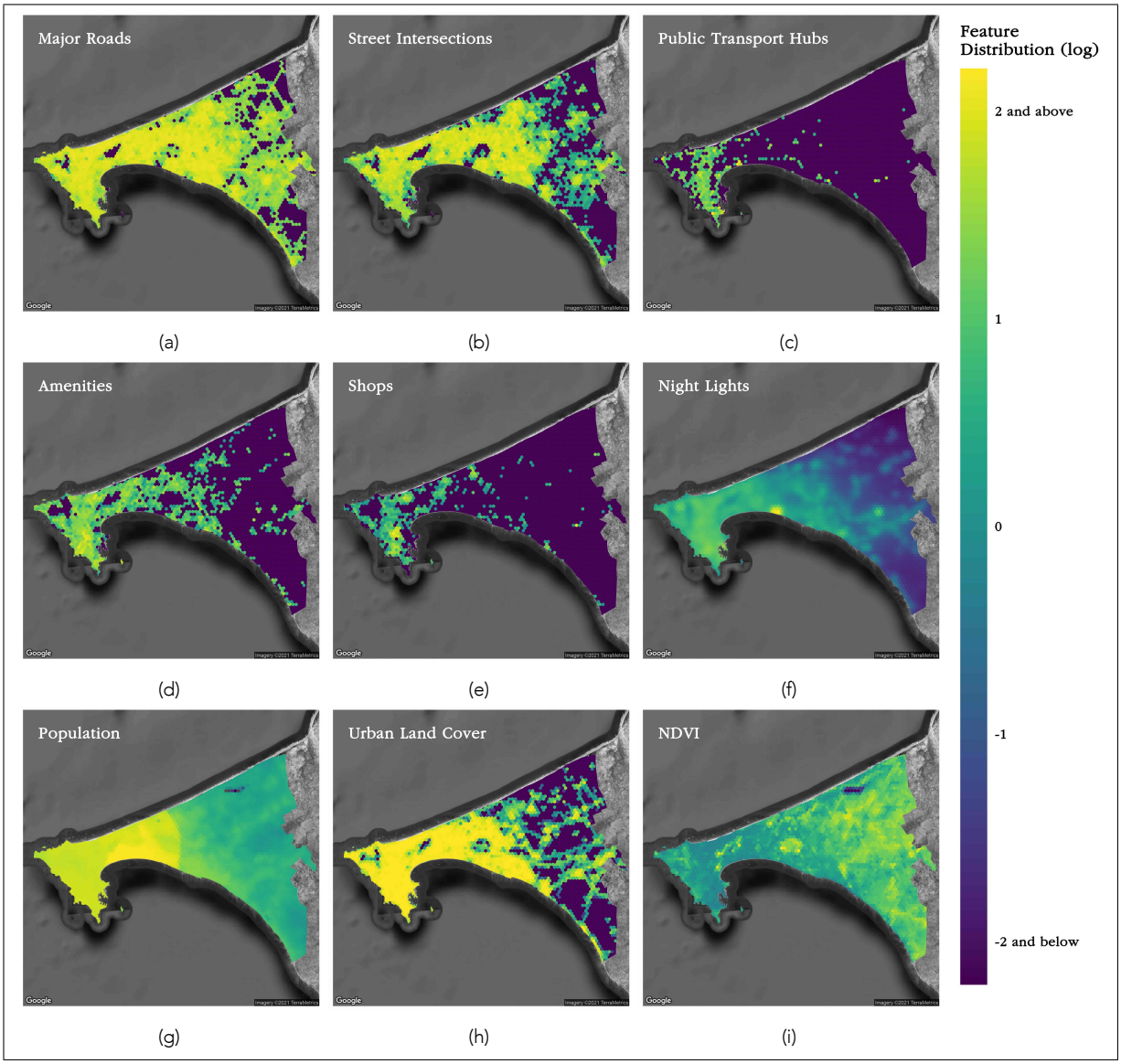

Figure 2: Illustration of Selected Features at 500m x 500m Hexagonal Grid Cells.
Subfigures (a) to (e) are density adjusted; NDVI is not logarithm adjusted; all features are within-city scaled.

area) but remain very concentrated along the coast. They are found in the vicinity of the Port and along the southern coastline. The rest of the territory is dominated by residential areas that continue to expand to the East of the peninsula. Some features, e.g. major roads, exhibit a high density towards the west of the city, but are also present towards the east, whereas other features, e.g. public transport hubs, appear to be nearly absent towards the city's east. Overall, the distribution of employment within the city exhibits relatively strong spatial clustering with a Moran's I value of 0.91. Some of the features similarly reveal strong spatial clustering patterns, e.g., night lights, while others are more spatially spread across the city, e.g. public transport hub locations[4]. Given that spatially disaggregate data, i.e. features

[4]The Moran's I statistic (Moran (1948) and Anselin (1996)) allows to quantitatively describe the level of spatial

in lines, points and polygons, is aggregated while data obtained via larger polygons or grid cells is disaggregated to the grid cell level, i.e. air pollution data comes in 1km^2 resolution, higher spatial autocorrelation is present within the latter. Hence, the calculated Moran's I statistics are influenced both by differences in spatial clustering of the features, but also by the spatial resolution of the original data. In order to reduce the potential impact of the different levels of spatial (dis)aggregation on the subsequent analyses, we train our models on a large set of cartographic, geographical and geological features for the predictions of spatial employment within cities.

We rely on feature engineering to generate and trial multiple feature variations derived from the original raw data. In other words, we create variations of the original data and include them as additional features. These spatial density adjusted parameters (i.e., neighboring cell values), binary features derived from binned quartiles of the features' distributions across values, binary features indicating the presence or absence of a given feature in a grid cell, and count and logarithm adjusted versions. Our assembled data set[5] comprising all sample cities contains 98,594 observations[6] and more than 100 features[7], where each unit of observation is a 500m x 500m hexagonal grid cell. Given that cities differ in their physical extent, and due to a higher presence of original polygons with 'NA' employment values in some cities, the contributions of cities to the overall sample is not uniform (see Table 1 for details).

3 Methods

3.1 Dimensionality Reduction via Principal Component Analysis (PCA)

Identification of patterns within large data sets with many features is often difficult, e.g. due to the presence of collinearities. Methods of Unsupervised Machine Learning are predominantly employed to identify clusters of observations sharing similarities or to compress high dimensional data to reveal patterns within the data more clearly. Within this work, we specifically rely on Principal Component Analysis (PCA) for this purpose (Pearson, 1901). PCA identifies a finite number of linear combinations containing all features (Principal Compo-

clustering of a variable, where Moran's I values generally range from -1 to 1. A Moran's I of approximately 0 indicates a completely random spatial distribution, negative values point to spatial dispersion and positive values to spatial clustering of a variable. Moran's I values of visualized variables: Employment Density (log): 0.91***, Major Roads (log): 0.57***, Intersections (log): 0.66***, Public Transport Hubs (log): 0.51***, Amenities (log): 0.62***, Shops (log): 0.59***, Night Lights (log): 0.99***, Population (log) : 0.98***, NDVI: 0.74***, Urban Land Cover: 0.76*** (with *** representing $p < 0.01$, ** indicating $p < 0.05$ and * pointing to $p < 0.1$ significance levels computed via Monte Carlo simulations across 500 runs respectively; all variables are analyzed post within-city standardization)

[5] Data set and feature space are used interchangeably throughout this work

[6] This only includes observations with non-missing information.

[7] Features are reshaped according to each model's requirement resulting in slightly different numbers of variables across models.

nents/PCs) while preserving the covariance structure of the original feature space. Each PC consists of a linear combination of all features, where the contribution of each feature within a PC is determined by its assigned feature loading, which are higher for variables more important to the variance of the overall data set. Hence, for each observation of the original data a projected point can be derived via the PCs computed. However, given that PCA is an unsupervised dimensionality reduction technique it does not allow to explicitly focus on a response variable. Additionally, PCAs are limited due to their linearity assumption of the underlying data. Thus, for this work, we limit our use of PCA to exploring and visualizing the data.

3.2 Ensemble Methods: Random Forest and Gradient Boosting Machine Algorithms

Models based on decision trees offer an alternative methodological approach of computations for prediction problems. Decision trees split the full data into rectangular regions ('leaves') through covariates and threshold selection to minimize the average squared error. Predictions are then formed on each partitioned data subsample and averaged over all to obtain a prediction model for the full data set (Breiman et al., 1984). In order to avoid overfitting, and thus reduced out-of-sample performance, a penalty term is identified to determine optimal hyperparameters, e.g. tree depth, along the trade-off of the model's complexity and out-of-sample performance[8].

While a model based on a single tree offers the advantage that it can be understood relatively intuitively, combined algorithm models generally outperform those derived from a single algorithm. Models building combinations of many decision trees ('Ensemble tree methods') have thus emerged as reliable extensions to single decision trees delivering robust performance quality. The Random Forest (RF) algorithm (Breiman, 2001) has been established as one of the most widely used ensemble tree methods across the recent literature (see for example Engstrom et al. (2021)). The RF algorithm generates mean predictions based on many randomly perturbed parallel trees where each tree is identified for a subsample of the overall data, where subsamples are most often identified via bootstrapping ('bagging'). Different from traditional decision tree models, RF models optimize their parameters for a random subset of covariates whose selection differs at each split. This introduces variation within the analyzed data for each tree along both the data splits and covariate selection and results in a better overall fit of the combined algorithm. In particular, the RF specific partitioning of observations and feature space rules allow for the model to remain robust even in the context of noisy data (Perlich et al., 2003). Additionally, given that the final RF model is based on averaging across many trees, RF models also smooth out any discontinuities introduced into the feature space through subsampling. RF generally outperform single decision tree based models, deliver robust high predictive performance while requiring little

[8] See Athey and Imbens (2019) for more details.

additional manual tuning. Gradient Boosting Machines (GBM) are a related ensemble tree method (Friedman, 2002). GBM also merge multiple tree models but different from RF estimate these sequentially and combine them by reiteration. Unlike RF, GBM models rely on 'boosting' to generate the final model, i.e. each tree is fit to the whole dataset and trees are added sequentially to enhance performance cumulatively. GBM models can deliver precise predictions, however they are noticeably more sensitive to outliers in the data and hyperparameters specifics than RF models (Dietterich, 2000).

Overall, both RF and GBM present algorithms that can deliver high precision estimates. However, while GBM can outperform RF under specific data structures, it is also significantly more sensitive to noise in the data and to the choice of hyperparameters.

3.3 Feature Selection Approaches with Regularized GLM Models

Large feature spaces allow to explore the relationship of the dependent variable with various covariates. However, a large number of features may also introduce higher levels of collinearity across the feature space and may result in overfitting, thus lowering out-of-sample performance (Athey and Imbens, 2015). In order to reduce the aforementioned issues while retaining as much as possible of the information contained within the feature space, regularized linear models[9] amend linear estimations by adding feature selection identification to the overall estimation. In the first step, the regularized GLM model selects important features across the whole feature space according to

$$\beta_{reg} = \arg\min\Big\{ \sum_{i=1}^{N} \Big(y_i - \sum_{j=1}^{K} x_{ij}\beta_j \Big)^2 + \lambda \sum_{j=1}^{K} \beta_j \Big\} \tag{1}$$

where $\sum_{i=1}^{N}(y_i - \sum_{j=1}^{K} x_{ij}\beta_j)^2$ are the estimation's residuals, and λ is the penalty term applied across a feature space with N observations and K features. The regularization parameter can either be set by the researcher or determined through the model via cross-validation to minimize the model's RMSE. However, given that the regularization estimation also introduces a penalty on the important features (Varian, 2014), the regularization estimation is followed by a standard linear model estimated over the subset of selected nonzero variables. The most known versions of regularization models are Lasso (Tibshirani (1996); Belloni and Chernozhukov (2013)), Ridge (Hoerl and Kennard, 1970) and Elastic Net (Zou and Hastie, 2005) models where the models differ in the specifics of the penalty parameter applied. Lasso models use absolute value penalties to shrink the coefficients of unimportant variables to zero and thus tend to perform better for models with a finite set of important covariates. Ridge models rely on quadratic penalties to reduce the coefficients of unimportant variables, but unlike Lasso models never shrink these to absolute zero. Ridge estimations are hence

[9] Also referred to as penalized (linear) models throughout the literature

more suited to models where multiple variables exhibit similar importance. Elastic Net models combine elements of both Lasso and Ridge models. Regularized models can easily be used in combination with the standard econometrics models, and thus remains intuitive in its interpretation.

3.4 Model Setup

In order to identify the model that best predicts employment across all sample cities, we test the algorithms outlined in Sections 3.2 and 3.3 across the whole data set and on separate SSA and LAT subsets. For all models, we follow a standard test-and-train split of the data. We generate a random 80/20 split where we train the models on 80% of the original feature space and test on the held back 20% to gain insight into the model's out-of-sample predictive performance; the test and train sets are identical across the models. For the GBM of Section 5.1 and Reg GLM model of Section 5.2 we further impose a k-fold cross validation (CV) set up with 10 folds and 10 repeats. Within the training of the models, the training data is thus split into 10 equally sized disjoint subsets where a model is fitted on 9 folds and internally assessed on the remaining fold. This is iterated over all folds and repeated 10 times, so that each model effectively runs through 100 fitting processes. This set-up enables us to identify and subsequently incorporate model hyperparameters that optimize out-of-sample performance. The final model is then fit to the complete train data set. Given the bootstrapping structure of random forest algorithms, no k-fold CV is required for hyperparameter optimization. To obtain insight into out-of-sample performance, we test the model on the test data and compare the models' performance by relying on the model's R^2 and Root Mean Squared Error (RMSE)[10]. R^2 presents how well the model predicts (the variation in) the response variable and the RMSE measures the averaged difference predicted and observed response value.

4 Exploratory Analysis

In order to gain insight into the structure of the data, we exploit two options to summarize the data: (1) Pearson correlations to identify pairwise relationships between all variables, and (2) Principal Component Analysis (PCA) to gain a thorough insight into the underlying dynamics of the feature space.

Figure 3(a) provides a visual representation of the pairwise Pearson correlations across the main features of the data set, and Figure 3(b) specifically depicts the pairwise Pearson correlations of the response variable of log(employment density) with the covariates at the grid

[10] $R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$ and RMSE = $\sqrt{\frac{1}{n}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$ where y_i is the observed, $\hat{y}_i$ is the predicted, $\bar{y}_i$ is the mean response variable, n is number of observations with $0 \leq R^2 \leq 1$ and RMSE is measured in the unit of the response variable.

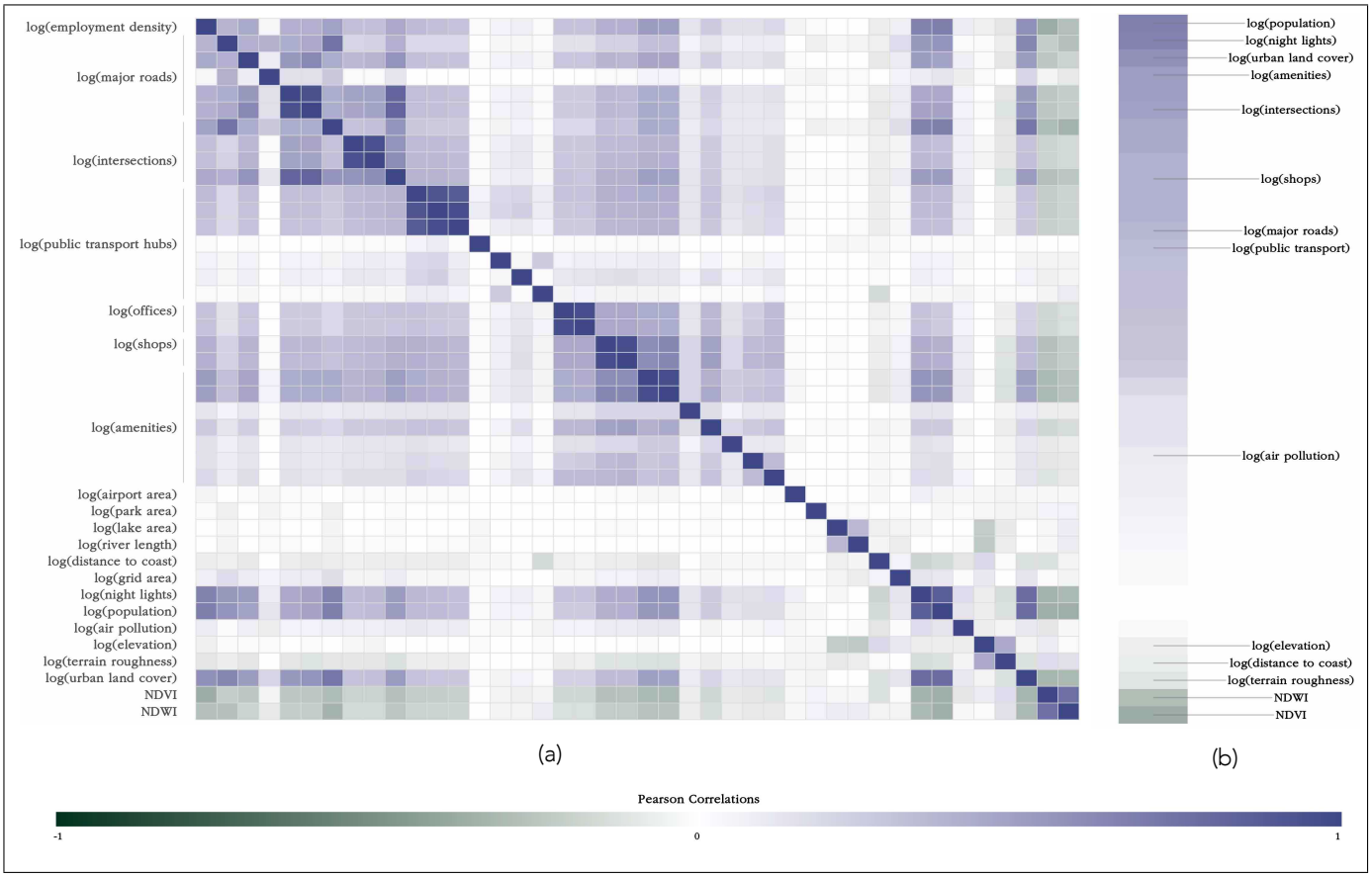

Figure 3: Overview of the Feature Space Correlations.
Subfigure (a) visualizes the pairwise Pearson correlations across the whole data set, and (b) summarizes the pairwise Pearson correlations of log(employment density) with the main features.

cell level across the whole data set. We identify correlations along both directions, i.e. there exist both positive and negative correlations across features, and with the response variable. Overall, the majority of correlations across the feature space are positive, and the positive strongly outweigh negative correlations in magnitude. We observe that the features derived from night lights and population data are most strongly correlated with employment. This is not unexpected given that information derived from night lights in particular have become a standard proxy for economic development across the literature, although predominantly in the context of cross-country studies or studies conducted at larger spatial units (see Henderson et al. (2012) for example).

From the cartographic features derived via OSM, we find that amenities and street intersections are particularly strongly correlated with employment. This could possibly point to co-location of amenities with urban employment or that amenities themselves are the place of employment in the case of the former (Ahlfeldt et al., 2015). The latter indicates that the street network is denser around high employment areas (Gudmundsson and Mohajeri, 2013). We further observe that the geographical and geological features concerning vegetation in-

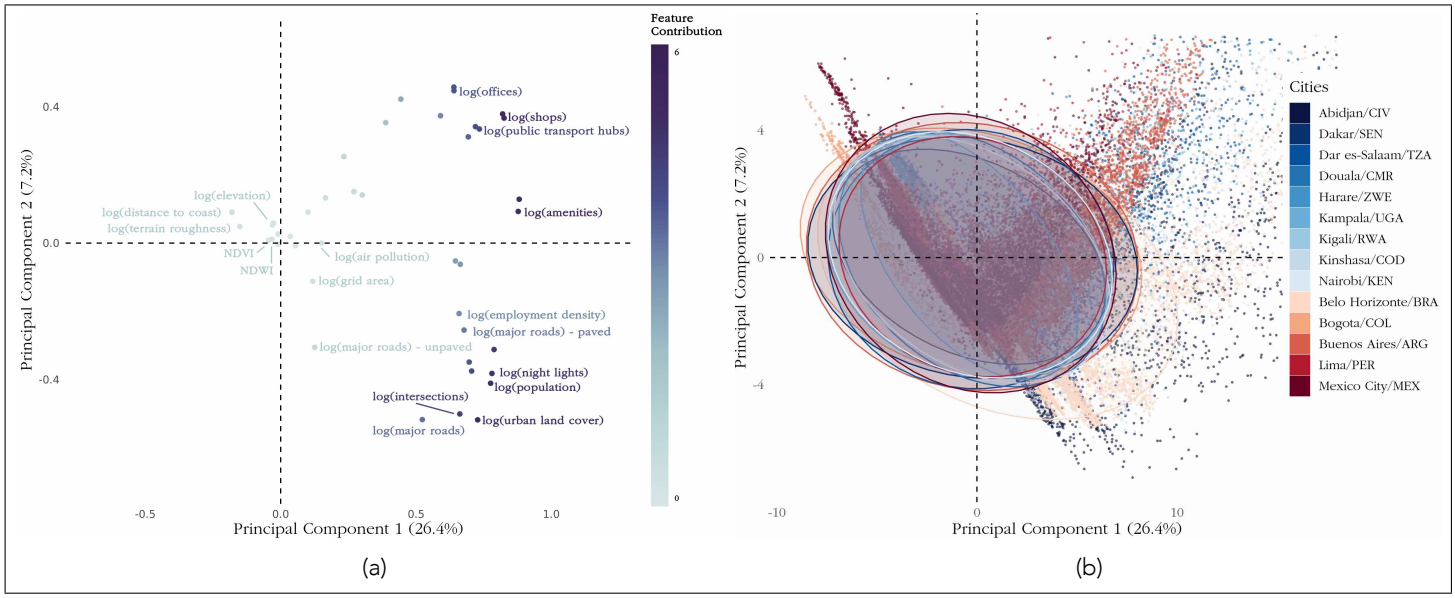

Figure 4: Principal Component Analysis where (a) is the Factor Loading Plot visualizing feature importance and (b) the Monoplot of the observations across cities in SSA (blue) and LAT (red)

tensity and water presence negatively correlate the strongest with employment, i.e. a higher presence of either water or vegetation indicates lower or no employment (see also Goldblatt et al. (2020)). Further, terrain roughness and elevation also tend to correlate negatively with the response variable, although with lower magnitudes. These also correlate negatively with the night lights and population variables, cautiously hinting at a reduced suitability of rough terrain for both residential and commercial buildings. In general, we observe a large number of positive correlations across the whole features space, and in particular features that are strongly correlated with the response variable also tend to be correlated themselves, and exhibit similar correlation patterns with other features.

In order to gain a deeper understanding of the underlying structure of the data, we further exploit a Principle Component Analysis (PCA). However, given that a PCA is an unsupervised approach, it does not allow us to explicitly focus on employment. Thus, we limit our use of the PCA to exploring and visualizing the data, and rely on it for two specific reasons: (1) to analyze how homogeneous the structure of the data is across cities following within-city standardization, and (2) to understand how the features relate to each other.

A PCA constructs various linear combinations of the original observations by assigning weights to each variable[11]. A principal component refers to a specific combination of feature weights. Figure 4(a) shows the PCA loading plot which illustrates how strongly each variable influences principal components 1 and 2 respectively, i.e., the weight assigned by each of the principal components to a given variable. The position on the graph responds to the

[11] See Section 3.1 for methodological details

relative weight identified for the variable by principal component 1 on x-axis and by principal component 2 on the y-axis. Night lights for example have been assigned a weight of 0.78 by principal component 1, and -0.38 by principal component 2. Given that larger weights represent variables more important to the data set, variables located further away from the origin can therefore be interpreted as more influential; equivalently, variables located closely to the origin are less important. The PCA loadings plot highlights that data on night lights, population, urban land cover, shops and amenities strongly contribute to the first two principal components hinting at the importance of these variables to the overall data patterns across the whole feature space. Further, each feature's position within the graph enables us to gain further insight into the correlations across the feature space. Features whose coordinates are within close proximity are generally positively correlated, and those who are located at an approximately 90° angle from each other are weakly correlated or uncorrelated. Features that diverge, and are thus located at approximately 180° from each other are negatively correlated. We observe that employment is most closely related to night lights, population and paved major roads, and to urban land cover and intersections data to a slightly lesser degree. This indicates that these variables correlate strongly with employment. This finding is in line with the Pearson correlations computed previously. This further suggests that night lights, population and urban land cover are important to both the variance of the whole data set and to predict employment, where the former follows from the variable's own position and the latter from its close location to employment on the graph. Furthermore, those features that exhibit low levels of correlations with employment, e.g., terrain roughness and elevation, also contribute relatively little to the PCs; this is indicated by their close location to the origin. This indicates that these variables likely do not contribute strongly to the prediction of employment given their weak correlation with employment as indicated by the PCA.

Relying on this PCA weighting structure, each original observation can be transformed through weighting the observation's variable values accordingly. Figure 4(b) shows the PCA's monopolot where each point refers to a observation transformed into an index through weighting of its variable value according to different weighting structures computed; the position on the x-axis refers to the observation's values weighted according to principal component 1 and the position on the y-axis according to principal component 2. The ellipsoids capture the area where the majority of the transformed observations are located where each ellipsoid refers to the observations of a specific city. Given that the city ellipsoids overlap to a large degree, i.e. the transformed observations are located largely within the same area, we can conclude that these observations are largely similar post-transformation. This allows us to conclude that the cities can be modelled jointly within one algorithm. A small degree of heterogeneity across and within the cities' ellipsoids can be observed. This is most likely driven by differences in data collection, heterogeneity in the relationship between response variable and covariates, and city specific idiosyncrasy.

5 Results

5.1 Random Forest and Gradient Boosting Machine Algorithm Predictions

In order to select candidate models for employment prediction, we consider our problem and the data sets at hand. On one side, we have employment data that is measured very differently across the sample cities. On the other side, we have OSM data that may be incomplete to varying degrees across cities. Given this statistical noise, we expect that the random forest (RF) model will be most suited due to its feature space and observation partitioning setup which makes it more robust in the presence of noise. Table 2 contains the results.

Across the whole data set, RF achieves high predictive performance with R^2 of 0.82 and RMSE of 0.43 (Table 2, column (1)). This result is substantially above those obtained in the related literature estimating employment factors. For instance, Goldblatt et al. (2020) predict enterprise and employment counts for Vietnamese communes and obtain R^2s ranging from 0.24 to 0.33. Our results are above but more comparable to those obtained from studies predicting poverty parameters where R^2s generally range from 0.3 to 0.8 (Jean et al. (2016) and Yeh et al. (2020)). However, unlike our approach that is based on pre-processed satellite data and mapped features these studies require complex deep learning methods to achieve their predictive performance. We do not observe noticeable differences in predictive quality across the two geographic regions. However, to test if models trained on data from each region separately achieve higher predictive performance, we retrain the algorithm on the data split accordingly (Table 2, columns (2) and (3)). The results indicate a marginal increase in the predictive performance for the model trained on SSA data, but no change in predictive quality for LAT observations. However, both of these models perform noticeably worse when assessed on data from the other region. Hence, in the cases where grid cells from the city being predicted are not contained in the training set (i.e., Latin American cities predicted from a model trained only on SSA cities), the predictions tend to be lower.

Employment is often clustered in one area if the city is monocentrically organized or multiple areas for polycentric cities. These clusters are generally larger than our specific grid cell choice of 500m x 500m. We hypothesize that high employment grid cells are generally surrounded by other high employment grid cells thus forming a high employment cluster. Furthermore, given that our grid cell structure has been superimposed in an arbitrary manner, it may for example be that the public transport hub closest to a high employment location is located in a neighboring cell. Similarly, a grid cell exhibiting a relatively high number of public transport hubs with low or no number of public transport hubs in surrounding cells is unlikely to indicate an employment cluster. This highlights the importance of including information on surrounding cells into the analysis. We thus extend the feature space by adding feature values (e.g., mean values) from neighboring cells to mimic the structure of a spatial

Table 2: Performance comparison across ensemble tree models (R^2 (RMSE))

		Train Data					
		All Cities	SSA Cities	LAT Cities	All Cities	SSA Cities	LAT Cities
		(1)	(2)	(3)	(4)	(5)	(6)
		Random Forest (RF)			Spatial RF		
Test Data	All Cities	0.82 (0.43)	0.59 (0.65)	0.43 (0.80)	0.95 (0.23)	0.62 (0.63)	0.44 (0.79)
	SSA Cities	0.82 (0.43)	0.85 (0.40)	0.40 (0.83)	0.95 (0.23)	0.96 (0.21)	0.41 (0.82)
	LAT Cities	0.82 (0.43)	0.53 (0.71)	0.82 (0.42)	0.95 (0.24)	0.54 (0.70)	0.95 (0.23)
No. Obs.		98,594	57,017	41,577	98,594	57,017	41,577
		Gradient Boosting Machines (xgboost)			Spatial xgboost		
Test Data	All Cities	0.70 (0.55)	0.48 (0.77)	0.38 (0.86)	0.72 (0.53)	0.51 (0.77)	0.30 (0.89)
	SSA Cities	0.67 (0.57)	0.72 (0.53)	0.36 (0.89)	0.70 (0.55)	0.75 (0.50)	0.26 (0.92)
	LAT Cities	0.74 (0.51)	0.44 (0.83)	0.77 (0.48)	0.74 (0.51)	0.47 (0.83)	0.78 (0.47)
No. Obs.		98,594	57,017	41,577	98,594	57,017	41,577

Note: Test and train data sets used across the algorithms are identical; test data is subsampled to avoid repeated observations; No. Obs. refers to the full data sets of both train and test data for each sample;

lag model (Anselin, 1996) ('Spatial RF' in Table 2, columns (4) to (6)). Overall, the results improve noticeably in comparison to the models trained exclusively on within-grid cell feature values and achieve very high predictive quality, both when trained on the whole data but also when trained on regional subsets with R^2 of 0.95 and 0.96 across the three models. However, similar to the results of columns (1) to (3), while marginal improvements in the models are observed when trained on regional sub-samples, the model trained on regional samples exhibits noticeably lower predictive performance compared to the model trained on the combined data.

Given that Gradient Boosting Machine algorithms often out-perform random forest models, we retrain the algorithms accordingly as an alternative approach. Here we rely on the extreme gradient boosting machine specification (xgboost); Results are provided in Table 2. Across both versions of the feature space, the results obtained via xgboost remain below those achieved via RF. Similarly to the RF models, Spatial xgboost outperforms the non-spatial version. In contrast to the RF case, models trained on region samples result in a higher predictive performance for the LAT than SSA model. However, these results also remain below the predictive performance achieved via RF.

Comparing both algorithms, the data and variable partitioning structures of the RF algorithm make it less prone to overfitting and more robust in the presence of noisy data. While xgboost models may outperform RF in settings where data is less noisy (Dietterich, 2000), within the context of our feature space which exhibits structural noise stemming from city id-

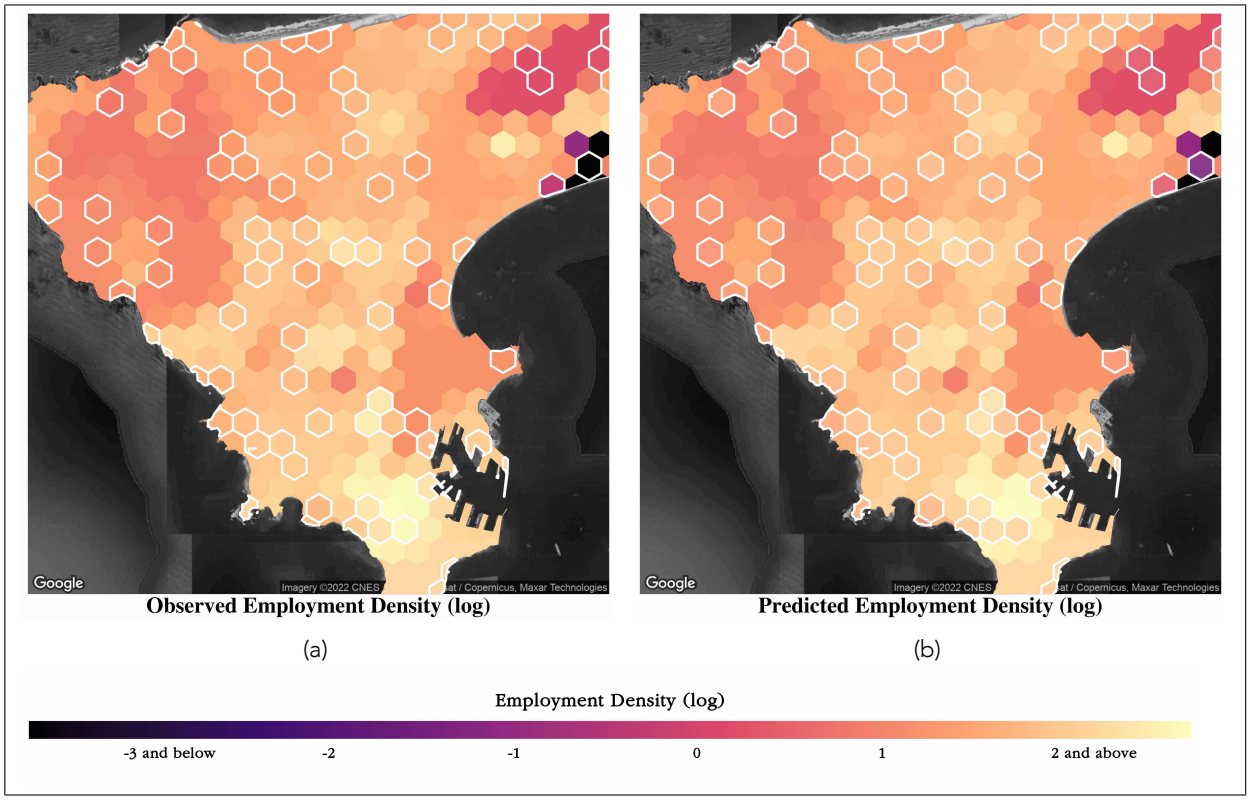

Figure 5: An illustration of predictions obtained via Spatial RF
(a) and (b) show observed and predicted values assigned of test data grid cells (outlined in white) within Dakar/Senegal

iosyncracy (city specific relationships between employment and features, and measurement errors in both the dependent variable and the OSM features), the RF model performs noticeably better across all variations of the data.

We can dig slightly deeper into model performance, and investigate how the model performs for small and large employment values. Typically, models tend to smooth towards the mean (performs the best for values close to the mean), while predicting less well outliers. An illustration of the observed and predicted grid cell employment values for Dakar, Senegal is provided in Figure 5. A visual comparison highlights an upwards bias for low levels of employment, e.g. the values predicted for grid cells located in the northeast of the area depicted. Some grid cells with high employment values appear to be underpredicted, e.g. in the south of Dakar, albeit to a lower degree.

Considering our full dataset, we find that higher values of employment are predicted more accurately but exhibit a small downwards bias while low employment levels tend to be slightly overpredicted (shown in Figure 6). This could be driven by repeated observations where polygons with zero employment have been disaggregated into multiple grid cells. Overall, despite minor smoothing, the relative squared error is roughly constant across dif-

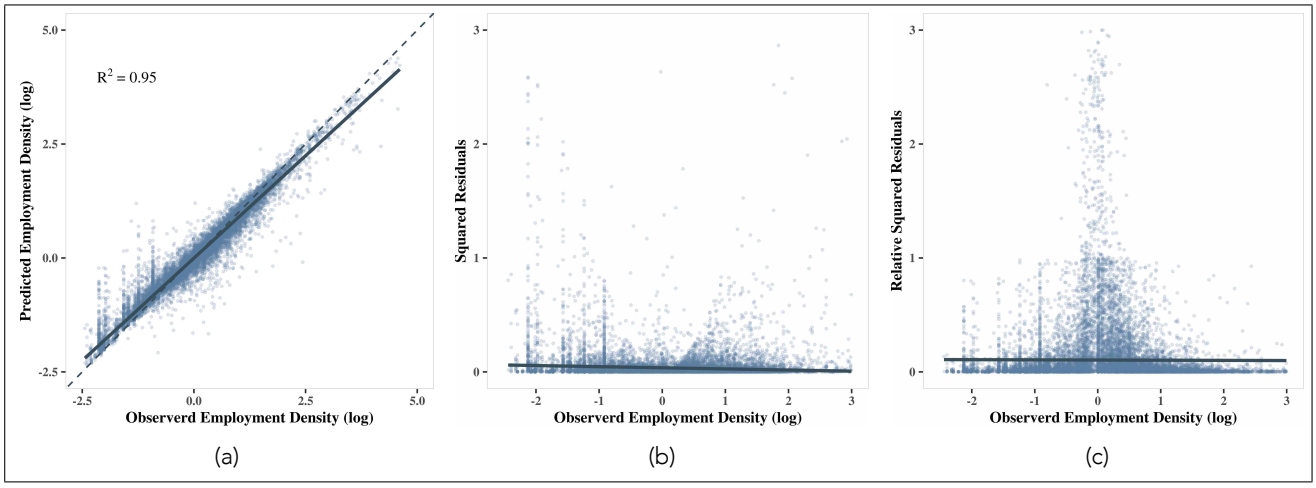

Figure 6: Performance metrics for Spatial RF

ferent levels of employment thus supporting the reliability of our results.

In order to gain an understanding of which features are particularly important for the prediction of employment, we extract the feature importance from the Spatial RF models trained on all cities and the SSA and LAT samples. A visualization of the most important features of each model is provided in Figure 7. For the model trained on all cities, population data (both cells and neighboring cells) is the strongest predictor of employment. This is followed by data on night lights and the NDVI observed across neighboring cells. The most important OSM-derived features are neighboring amenities (neighbor version of aggregated amenity variable) which are identified as the 5th most important feature. Noticeably, 11 of the 15 most important features are satellite derived. Further, features identified as important are often relevant in both own-cell and neighbor versions, where for example the population of a grid cell but also the population of neighboring cells are identified as important predictor variables. This might in part be driven by the spatial disaggregation employed when obtaining grid cell level values from polygons and might also be influenced by the non-random spatial distribution of both dependent and feature variables.

For the model trained solely on SSA cities' observations, the features of population, night lights and urban land cover are the most important. Interestingly, two specific features (intersections and public transport) are ranked as important for the SSA model but do not feature within the overall model. For the LAT cities' model, urban land cover is also highly important in both its within grid and within neighboring grid cells. Across all three models, three main conclusions regarding feature importance can be drawn: (1) satellite data derived features are usually most important, (2) amenities (a combination of both categorized and uncategorized amenities) appear to be the most relevant employment predictor derived via OSM, and (3) some features are identified as influential across all three models, but there exists also heterogeneity across the samples.

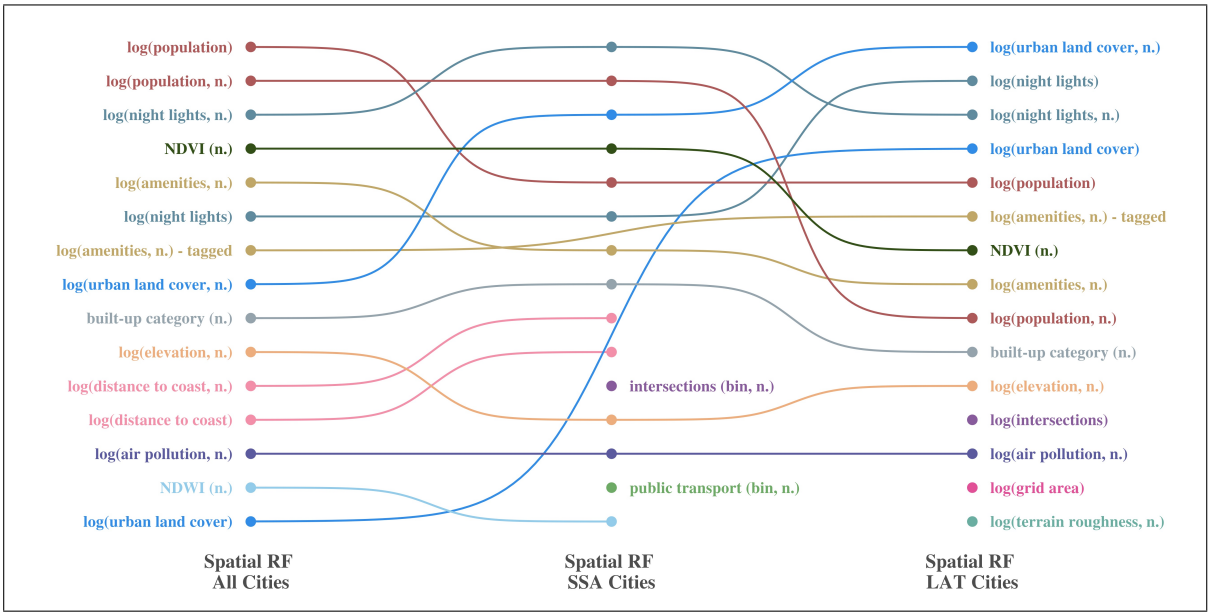

Figure 7: Variable importance ranking of the 15 most important features for Spatial RF

Overall our results reveal that a combination of satellite and mapped features data can predict employment to a high degree; even for the high levels of spatial resolution required for within-city analyses. Furthermore, despite the presence of noise in the data, exploiting the RF algorithm results in very high predictive performance quality. In addition, adding neighboring features delivers particularly high predictive performance, highlighting the importance of explicitly incorporating the geospatial dimension of data. Interestingly, we find that models trained on one region but tested on the other perform less well. This could be due to two distinct phenomena: first, there could be structural differences between the two sets. In other words, cities in LAT could share common relationships between employment and our features (but different to SSA). Second, cities could themselves exhibit unique relationships, and hence leaving own-city cells out of the training data significantly impacts the ability of the model to predict that city.

Another consideration that deserves some attention is that of heterogeneity in data quality across cities. Both the underlying surveyed employment data that we are trying to predict and the OSM completeness could be a main cause of differences in relationships between employment distribution and mapping and satellite features. Such data quality issues could be consistent with both hypotheses: structurally different relationships in SSA and LAT if the OSM data is consistently less complete in SSA cities; and differing prediction quality if employment data is of lower quality in particular urban areas. We will further investigate these issues below.

5.2 Alternative Algorithms

In order to test whether algorithms based on different statistical foundations provide better predictive performance, we fit regularized linear algorithms on the training data and benchmark their performance. Within the training process of RF models the data is randomly split and a random subset of variables is tested for each data subset. Given this evaluation of a randomly drawn subset of features, the likelihood that each tested subset contains highly correlated variables is low. This set-up reduces the issue of multicollinearity within RF models (Genuer et al., 2010). Regularized linear models do not partition the data, but instead first reduce the feature space to only the most important variables before fitting the model in the second step. As penalized models represent a family of models differing in their penalty specifics, we allow the algorithm to determine the most suitable model through optimal penalty parameter choice. Results are provided in Table 3.

Overall, the model selects the clear majority of features as relevant across both non-spatial and spatial versions. Comparing the predictive performance reveals that the models' R^2 generally remains substantially below (and RMSE above) those observed for both RF and GBM models. This holds true across all six versions of the fitted model (spatial vs non spatial, all cities, and regional specific). Similar to the GBM fit models, the algorithm fit exclusively on LAT cities tends to perform better than those fit to all cities' or SSA cities' data. However, the increased fit to the data is similarly accompanied by overfitting, i.e., the model is very closely aligned with the specific training data and thus not sufficiently flexible to predict any other data to a high degree. This results in worse predictive performance when the algorithm is used to predict the SSA or overall cities' test data. As with the ensemble tree simulations, the models based on data including spatial effects outperform the non-spatial version although increases in predictive performance are very marginal.

Hence, the predictive performance of these models is lower compared to the RF models discussed in the previous subsection. However, we can exploit the fact that regularized linear models select a subset of important features in order to delve deeper into understanding how different features interact with employment. For example, the model selects 81/84 features for the 'All Cities' model. Taking this subset in each case, we run a spatial regression analysis which directly incorporates spatial dependence within the data. We examine the model coefficients, comparing the most important features to those previously identified by the RF feature importance score; Table 3 provides the results[12]. Across all three regression sets, the features of NDWI and NDVI (water/vegetation) are reoccurring with high importance. Night lights are the only additional satellite derived feature listed for the 'All Cities' model, while terrain roughness, population and urban land cover are also listed for the 'SSA' and 'LAT'

[12]All Spatial Regressions listed in Table 3 follow the Spatial Durbin Error model; given the large amounts of features, we only list the largest ten coefficients among statistically significant variables.

Table 3: Performance comparison across Regularized Gen. Linear Models (R^2 (RMSE))

		Train Data					
		All Cities	SSA Cities	LAT Cities	All Cities	SSA Cities	LAT Cities
		(1)	(2)	(3)	(4)	(5)	(6)
		Regularized Generalized Linear Models (Reg GLM)			Spatial Reg GLM		
Test Data	All Cities	0.56 (0.66)	0.53 (0.73)	0.44 (0.75)	0.58 (0.65)	0.52 (0.78)	0.46 (0.74)
	SSA Cities	0.49 (0.72)	0.51 (0.69)	0.43 (0.76)	0.51 (0.70)	0.54 (0.68)	0.45 (0.75)
	LAT Cities	0.67 (0.58)	0.55 (0.73)	0.69 (0.56)	0.68 (0.57)	0.54 (0.81)	0.70 (0.55)
Variables Selected		81/84	75/84	69/84	128/142	129/142	126/142
No. Obs.		98,594	57,017	41,577	98,594	57,017	41,577

Spatial Regression Analysis		
All Cities	SSA Cities	LAT Cities
-0.88*** (0.05) NDVI	-1.05*** (0.21) Intersections (binary)	0.96*** (0.08) NDWI
0.78*** (0.06) NDWI	-0.94*** (0.08) NDVI	-0.87*** (0.07) NDVI
-0.48*** (0.13) Offices (binary)	0.79*** (0.29) log(shops)	-0.58*** (0.18) Major Roads (binary)
-0.36*** (0.02) Major Roads (binary)	0.37*** (0.08) NDWI	0.32*** (0.02) log(intersections)
0.33*** (0.06) Aiport (binary)	0.31*** (0.07) Shops (binary)	0.27*** (0.02) log(night lights)
0.28*** (0.01) log(night lights)	-0.29* (0.17) log(trams)	0.24*** (0.07) Offices (binary)
0.27* (0.15) Shops (binary)	0.26** (0.11) Airport (binary)	0.24*** (0.05) Amenities (binary)
0.27*** (0.02) log(major roads)	0.25*** (0.03) log(terrain roughness)	-0.18*** (0.07) log(urban land cover)
-0.24*** (0.02) Intersections (binary)	0.11*** (0.04) log(main roads)	0.16*** (0.02) log(population)
0.20*** (0.02) log(intersections)	0.10*** (0.02) log(population)	-0.15*** (0.04) log(intersections with major roads)

Note: Test and train data sets used across the algorithms are identical across the different algorithms; test data is sub-sampled to avoid repeated observations; No. Obs. refers to the full data sets of both train and test data for each sample; Estimated coefficients are reported for the ten most important variables in each regression as identified through the coefficients' absolute magnitude. A Spatial

models. Among OSM features, information on major roads and intersections are relevant for all three models, even though through different features; variables describing shops, airports, amenities and offices are also important, however to very differing magnitudes. Surprisingly, only one variable reflecting public transport is listed. While this is surprising given the literature on the importance of public transport for urban employment (Pogonyi et al., 2021), our results may be driven by the absence of well-established public transportation networks in the majority of the cities included here. Most of the variables identified here as important are equivalent to those exhibiting strong pairwise correlations with employment[13] and those selected as important by the random forest models in the preceding subsection. However, large differences in variables indicated as important by the linear and random forest models can also be observed. For example, information on air pollution is selected as important across all RF models, but it is not identified as such by the linear regressions. Additionally, the results of Table 3 identify binary variables as important while the random forests computations predominantly list continuous variables as highly relevant. These differences with the RF models are most likely driven by the less strict assumptions of RF models that allow for

[13]See Section 4; a direct comparison with the PCA results is not possible given that the PCA is an unsupervised approach that identifies variables importance to the overall data set where as the analyses of this section identify variables specifically important for employment.

Table 4: Performance Comparison across RF Models with Spatial Effects (R^2 (RMSE))
OSM vs. Satellite Features Only

		Train Data		
		All Cities	SSA Cities	LAT Cities
OSM Data Only				
Test Data	All Cities	0.76 (0.49)	0.55 (0.68)	0.39 (0.80)
	SSA Cities	0.76 (0.49)	0.78 (0.47)	0.36 (0.82)
	LAT Cities	0.75 (0.50)	0.50 (0.73)	0.76 (0.49)
No. Obs.		98,594	57,017	41,577
Satellite Data Only				
Test Data	All Cities	0.79 (0.47)	0.54 (0.69)	0.43 (0.80)
	SSA Cities	0.78 (0.48)	0.80 (0.45)	0.41 (0.82)
	LAT Cities	0.80 (0.45)	0.48 (0.75)	0.81 (0.45)
No. Obs.		98,594	57,017	41,577

Test and train data sets used across across the algorithms are identical; test data is subsampled to avoid repeated observations; No. Obs. refers to the full data sets of both train and test data for each sample;

the detection of variables as important even if their relationship to employment is non-linear. These results, similar to those of Figure 8, indicate a strong importance for satellite derived features, although RF models identify night lights and population data as most crucial while the regression estimate NDVI and NDWI variables as most important. In contrast to the RF feature importance, amenities are estimated to be less important by the regression analyses, while features derived via the street network, i.e. major roads and street intersections, are indicated as important for employment.

Overall, the high correspondence between the variables selected as important by the RF and linear models provides confidence that these variables are highly predictive of employment density.

5.3 Predicting Employment from Digital Maps

The results of the preceding section strongly suggest that urban employment can be predicted from a combination of mapped and satellite data. Variable importance analyses further highlighted the high relevance of satellite derived features to predict employment. In this section, we collapse the feature space including spatial effects to those variables derived via OSM to test how well employment can be predicted from mapped or satellite information alone. We retrain the RF with spatial effects algorithms on the reduced feature spaces.

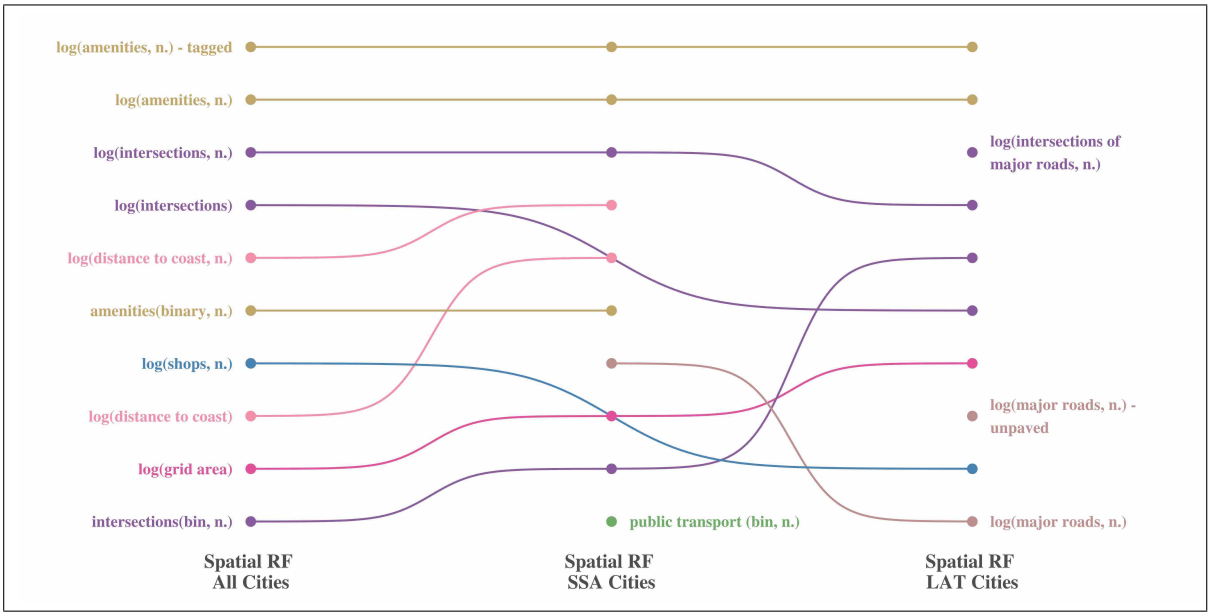

Figure 8: Variable importance: 15 most important OSM features for Spatial RF

Results are provided in Table 4.

Overall, the results indicate that a large part of the variations in employment can be predicted solely based on OSM data with an R^2 of approx 0.76, and an R^2 of approx 0.79 for satellite only data. In comparison to the previous computations relying on the combined data, omitting one or the other type of data presents a significant reduction in predictive performance. In the case of OSM, we might expect LAT cities to be better predicted due to better quality data - but this is not the case. This may be due to a large sample size in SSA. In the case of satellite data, we expect that similar data quality across LAT and SSA would yield similar predictive power. This is consistent with our findings.

Similar to the algorithms outlined in the previous sections, we investigate which variables are particularly important for predicting employment density. Results are provided in Figure 8. Similar to the previous results, amenities are indicated as the most important OSM derived feature. This holds true across all three algorithms, and refers particularly those amenities that exhibit a tag, i.e. amenities that can be identified as a restaurant, etc. Different features capturing street intersections are also indicated as particularly important, although in different variations across the three models. Features capturing the grid cell's distance to the coast and features indicating shops are also revealed to be of importance, although the former more so for SSA cities and the latter for LAT cities. Features representing major roads and public transport appear to be only important to a smaller degree. The lower importance of public transport could be due to relatively low public transport availability in the majority of cities included in this analysis.

5.4 Predicting Out-of-Sample Cities

Our ultimate aim is to be able to predict employment in entirely unseen or out-of-sample cities. Here, we use training data for 14 cities in order (a) to test the feasibility of this and (b) to rely on the algorithms to predict urban employment of cities not included in this sample based on the data patterns identified.

If we randomly split the data into training and test data sets, then grid cells from any given city appear in both sets. In this case we do not know how well the algorithm would perform on a city that was never included in the training sample. In the analyses above, algorithms fit on SSA and LAT samples point to reduced accuracy in prediction when evaluated on cities located in the other region. Hence, city idiosyncrasy with regards to both noise in the data and/or the relationship between employment and features can substantially decrease the algorithm's predictive performance for cities fully excluded from the training data.

We can investigate the predictive power of the model for unseen cities by refitting the spatial RF model to the complete data set comprising all grid cells but excluding a specific city. Hence, for each city, we fit our algorithm to the grid cells of the other thirteen cities and evaluate the model performance on the grid cells of the city held back. The raw employment data was provided in polygons. Given that spatially disaggregating polygon information into grid cells might introduce noise to the data, we evaluate the predictive performance of the algorithms both at the grid cell level, i.e. the level that we use to train the algorithms, and also spatially aggregated to original polygon boundaries. Results are provided in Table 5. Overall, when evaluated at the grid cell level (column 1) our results exhibit large heterogeneity in predictive performance with the lowest R^2 of 0.20 for Harare/ZWE to the highest of 0.71 for Buenos Aires/ARG. Figure 9 visualizes observed and predicted employment for these two cities. While these results are overall noticeably lower than those of section 5.1 to 5.3, the predictive performances are in line with a similar analysis undertaken to predict consumption expenditure and household assets of spatial clusters across four SSA countries (Jean et al., 2016). Our algorithms predict employment at a grid level based on detailed input data. When we aggregate up to polygon level, we would expect that our employment predictions are well-aligned with the observed employment data (which was collected at polygon level). Indeed, we do find that we obtain substantially higher values of R^2 for almost all cities when evaluated at polygon level (column 2).

In columns (4) to (6) we retrain an algorithm for each city, but only train on grid cells from other cities located in the same region. The similarity of predictive performance of the models fitted on all cities and those fitted solely on cities of the same geographical region does not point to a clear better fit from either approach. Although some cities are predicted more accurately when the regional models are used, no clear overall pattern emerges supporting an obvious choice for either. Similar to the results of column (2), evaluating the SSA and LAT

Table 5: Performance Comparison across RF Models with Spatial Effects for Out-of-Sample Cities (R^2 (RMSE))

	Model (unit of evaluation)					
City	All Cities (grid cell)	All Cities (polygon)	SSA Cities (grid cell)	SSA Cities (polygon)	LAT Cities (grid cell)	LAT Cities (polygon)
SSA						
Abidjan/CIV	0.35 (0.84)	0.70 (1.25)	0.30 (0.85)	0.71 (0.33)	-	-
Dakar/SEN	0.62 (0.62)	0.70 (0.48)	0.65 (0.59)	0.68 (0.48)	-	-
Dar es Salaam/TZA	0.40 (0.78)	0.71 (0.75)	0.40 (0.77)	0.71 (0.74)	-	-
Douala/CMR	0.37 (0.81)	0.65 (0.56)	0.40 (0.79)	0.64 (0.56)	-	-
Harare/ZWE	0.20 (0.91)	0.30 (0.48)	0.17 (0.93)	0.28 (0.48)	-	-
Kampala/UGA	0.54 (0.68)	0.54 (0.68)	0.49 (0.71)	0.49 (0.71)	-	-
Kigali/RWA	0.52 (0.71)	0.81 (0.72)	0.49 (0.73)	0.79 (0.72)	-	-
Kinshasa/COD	0.49 (0.72)	0.55 (0.55)	0.39 (0.79)	0.49 (0.69)	-	-
Nairobi/KEN	0.67 (0.61)	0.77 (0.62)	0.67 (0.63)	0.80 (0.58)	-	-
LAT						
Belo Horizonte/BRA	0.62 (0.82)	0.77 (0.90)	-	-	0.64 (0.61)	0.84 (0.40)
Bogotá/COL	0.58 (0.65)	0.52 (0.45)	-	-	0.58 (0.65)	0.53 (0.45)
Buenos Aires/ARG	0.71 (0.54)	0.78 (0.28)	-	-	0.71 (0.56)	0.79 (0.32)
Lima/PER	0.50 (0.72)	0.42 (0.50)	-	-	0.48 (0.73)	0.42 (0.53)
Mexico City/MEX	0.58 (0.67)	0.58 (0.67)	-	-	0.56 (0.67)	0.56 (0.67)

Note: Original employment data for Kampala and Mexico City was provided as point locations, which were aggregated to grid cell levels for computations; for these cities all performance evaluations were conducted at grid cell levels

models at polygon levels results in substantially higher predictive performance levels.

We hypothesize that the different sizes of polygons across cities, and associated levels of aggregation, influence the performance of the algorithms. For example, in cases where the observed employment data is collected as large polygons, we have a large mis-match between the level of detail of the predicted employment at grid level and the observed employment transformed to grid level. Hence, we would expect that cities with larger polygons are worse predicted at a grid level, while cities with small polygons are more accurately predicted. In order to investigate this, we plot each city's mean polygon size against its model R^2 (for grid cells) in Figure 10(a). We can see a clear negative relationship between these two metrics supporting this hypothesis. When we plot this relationship, but with both predicted and observed data aggregated to polygon level, we no longer observe a negative slope, see

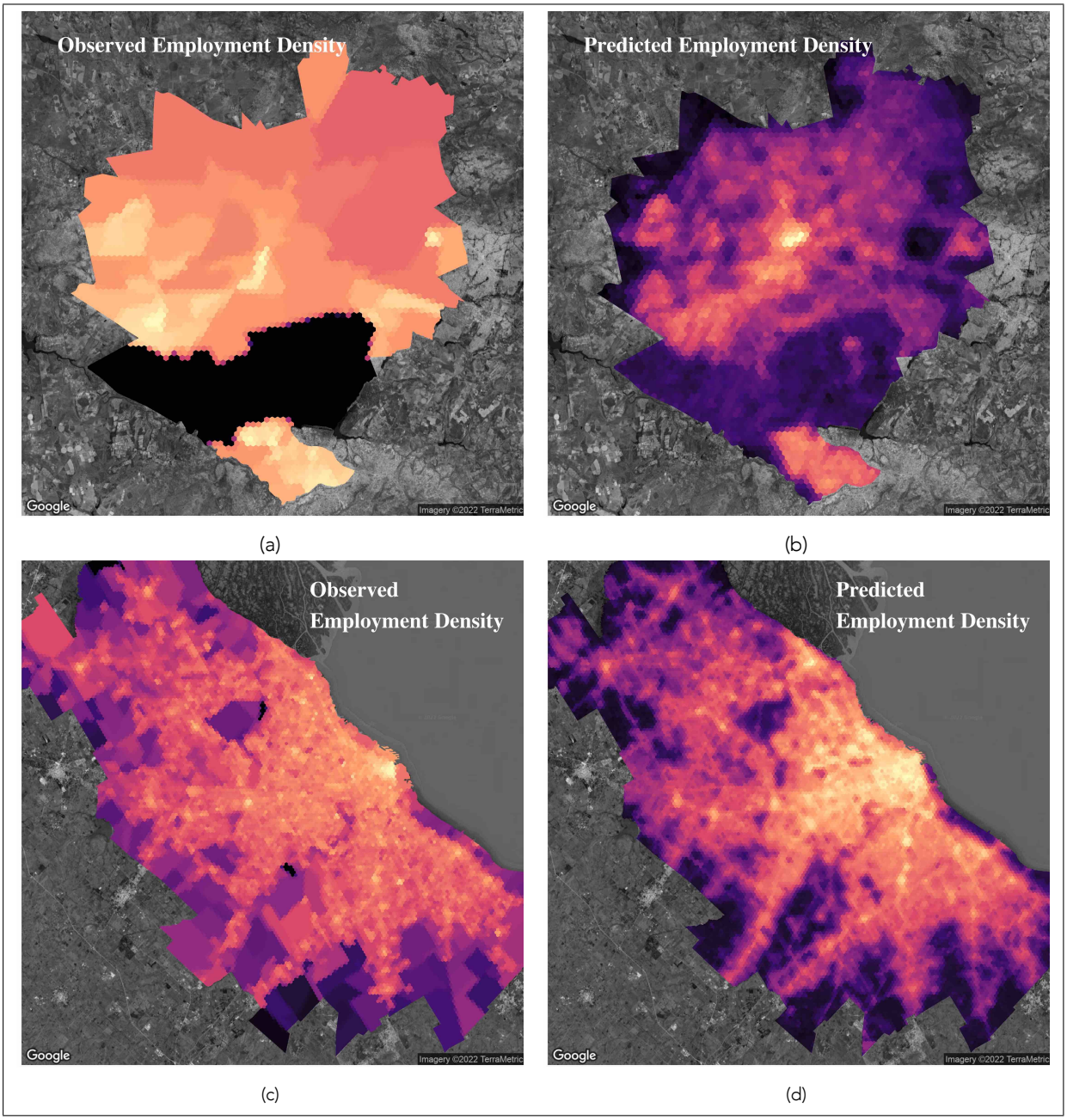

Figure 9: Observed vs. Predicted Employment for Harare/ZWE in (a) and (b), and for Buenos Aires/ARG in (c) and (d)[14]

Figure 10(b). Hence, we can deduce that the quality of our predictions is independent of polygon size.

To understand our models better, we analyze further patterns within our employment predictions. Across all cities, the results (a) exhibit smoothing towards the mean, (b) the upwards bias in low employment grid cells is on average larger than the downwards bias in high

[14] Predictions obtained via the models trained on all other cities

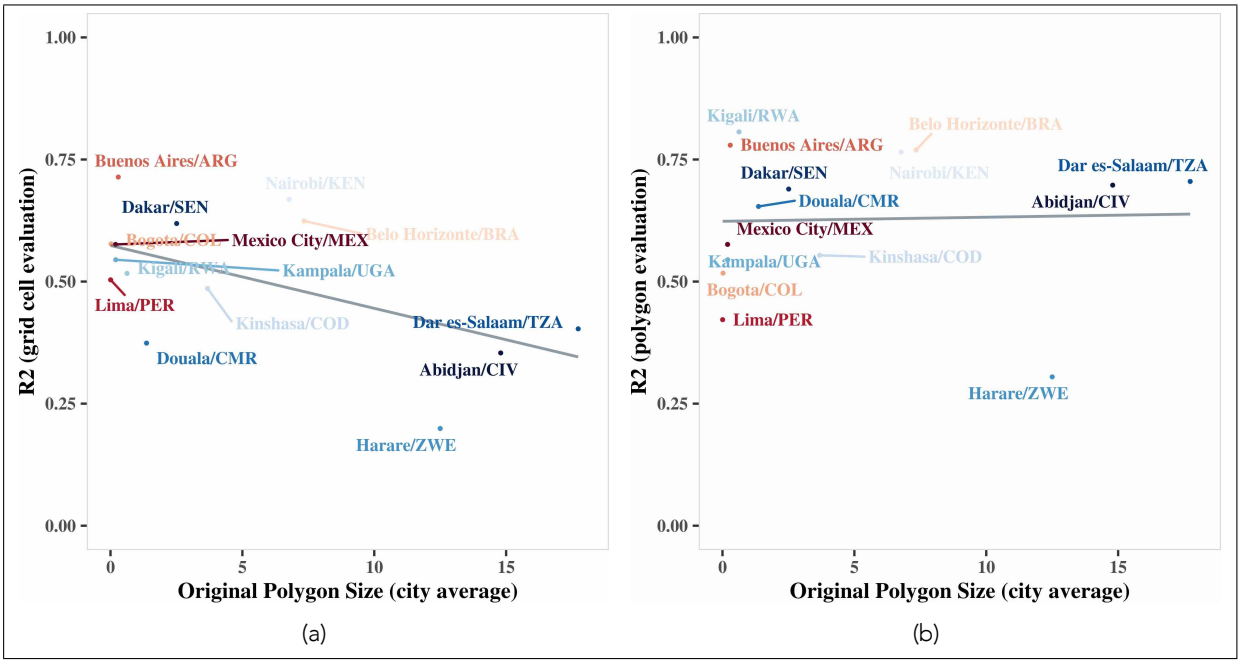

Figure 10: Correlation between R^2 and Average Polygon Size across Cities

employment grid cells, and (c) predicted employment is on average within 0.71 (absolute deviation) of observed employment (which, when normalized, ranges from about -3 to +3). The latter allows us to conclude that while there are differences between predicted and observed employment, on average a predicted value is within -/+ 0.71 of the observed employment within that grid cell. Hence, on average we do not see very large differences between observed and predicted employment. However, given the noise in the data, it is not possible to conclude if this is due to measurement noise or algorithmic performance.

In order to probe the degree of city idiosyncrasy across the cities, we train a unique model for each individual city. If a city can be described by its 'own' model, then we can deduce that there is no inherent limit on an algorithm to predict employment from the data on that city. Hence, we refit city specific models to 80% of each city's grid cells and evaluate on the remaining 20%. The R^2 across these models range from 0.68 for Lima to 0.99 for Harare with an average of 0.84. These results suggest that cities' employment can overall be predicted accurately. However, the results also show a significant amount of heterogeneity within feature importance across the models (data not shown). Hence, while urban employment can be predicted, city idiosyncrasy stemming from measurement noise in both employment and features, and heterogeneity in the relationship between employment and features, limits out-of-sample performance.

In summary, OSM and satellite data can predict employment to a very high degree. However, the accuracy of the predictions is influenced by both (a) a structurally different relationship between employment and included variables across the different cities and (b)

differences in OSM and employment data quality across the cities.

Since our guiding objective was to design an algorithm to predict employment in unseen cities, we apply it to various cities for which we have no employment data, Niamey/Niger, Khartoum/Sudan, Mumbai/India, Karachi/Pakistan, Port au Prince/Haiti and Guayaquil/Ecuador are provided as examples in Figure 11. As our results do not point to a clear preference for a model trained on all cities or trained solely on those cities of a given geographical region, we deploy the algorithm trained an all cities for better cross city comparison patterns here.

6 Discussion & Conclusion

We set out to tackle a key gap in the policy toolbox, the lack of highly granular data on employment within cities, particularly in less developed areas of the world. Using a spatial adaption of the random forest algorithm, we show that we can predict within-city cells in our test cities with extremely high accuracy (>95% R^2), and cells in out-of-sample cities with medium to high accuracy (21% - 72% R^2 at grid level and 31% - 80% at polygon level). While we found that our model picked up significant city-specific relationships, the moderate to high R^2 obtained for out-of-sample predictions, particularly for cities with expected higher data quality, gives us confidence that the algorithm can be deployed on unseen cities.

Our contribution to the literature and practical toolbox spans multiple levels. First, we show that a combination of mapped features and pre-processed satellite data can be used to fill data gap in developing countries. Importantly, we illustrate this for the tricky within city case, generating very high resolution estimates. Distinct from most other related studies (Jean et al. (2016) and Yeh et al. (2020)), we focus on employment as our outcome variable, rather than less precise metrics such as poverty or GDP. We find that our model captures a significant level of city heterogeneity. This likely comes from multiple sources, including both true heterogeneity in terms of the relationship between employment and our features, but also city-specific effects stemming from noise in both the employment data and the OSM data.

One of the key motivations for this study was to develop an algorithm that could be deployed on completely unseen cities, i.e., cities for which no employment/training data is available. Despite the presence of city heterogeneity in our data, we find an acceptable level of R^2 in the out-of-sample case, particularly for cities with expected higher data quality. This gives us confidence that, despite a few outliers with lower R^2, the model performs sufficiently well. Given the model testing above, we recommend deploying the spatial RF trained on all 14 cities as the base model to give the best prediction.

[15] Predictions obtained via the models trained on all other cities.

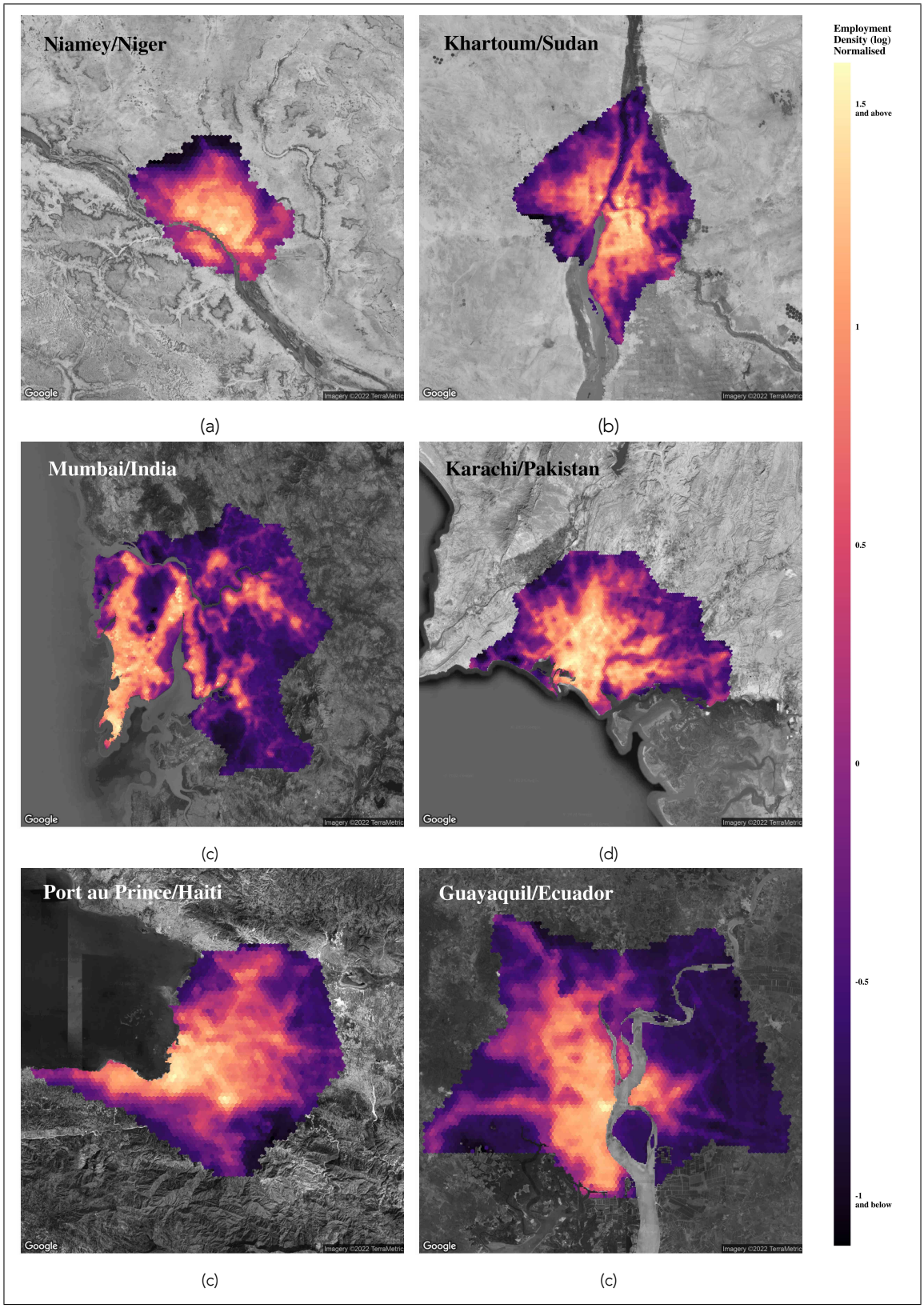

Figure 11: Employment Predictions across various Cities[15]

References

Ahlfeldt, G.M., Albers, T.N.H., Behrens, K., 2020. Prime Locations. CEP Discussion Paper No. 1725. Centre for Economic Performance/CEP.

Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2015. The economics of density: Evidence from the Berlin wall. Econometrica 83, 2127–2189.

Anselin, L., 1996. The Moran Scatterplot as an ESDA Tool to Assess Local Instability in Spatial Association, in: Fischer, M., Scholten, H., Unwin, D. (Eds.), Spatial Analytical Perspectives on Gis in Environmental and Socio-Economic Sciences. Taylor; Francis, London, pp. 111 – 125.

Athey, S., Imbens, G.W., 2015. Machine Learning Methods for Estimating Heterogeneous Causal Effects. Working Paper No. 3350. Stanford Graduate School of Business, Stanford University.

Athey, S., Imbens, G.W., 2019. Machine Learning Methods That Economists Should Know About. Annual Review of Economics 11, 685–725.

Baruah, N.G., Henderson, J.V., Peng, C., 2021. Colonial legacies: Shaping African cities. Journal of Economic Geography 21, 29–65.

Belloni, A., Chernozhukov, V., 2013. Least squares after model selection in high-dimensional sparse models. Bernoulli 19.

Bengtsson, L., Gaudart, J., Lu, X., Moore, S., Wetter, E., Sallah, K., Rebaudet, S., Piarroux, R., 2015. Using Mobile Phone Data to Predict the Spatial Spread of Cholera. Scientific Reports 5. URL: `http://www.nature.com/articles/srep08923`, doi:`10.1038/srep08923`.

Bertaud, A., 2002. The spatial organization of cities: Deliberate outcome or unforeseen consequence? World Development Report 2003, background paper URL: `http://siteresources.worldbank.org/DEC/Resources/spatialorgcity.pdf`.

Blumenstock, J., Cadamuro, G., On, R., 2015. Predicting poverty and wealth from mobile phone metadata. Science 350, 1073–1076.

Breiman, L., 2001. Random Forests. Machine Learning 45, 5 –32.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. Wadsworth & Brooks, Pacific Grove.

Bryan, G., Glaeser, E., Tsivanidis, N., 2020. Cities in the Developing World. Annual Review of Economics 12, 273–297.

Buchhorn, M., Lesiv, M., Tsendbazar, N.E., Herold, M., Bertels, L., Smets, B., 2020. Copernicus Global Land Cover Layers—Collection 2. Remote Sensing 12, 1044.

Burke, M., Driscoll, A., Lobell, D.B., Ermon, S., 2021. Using satellite imagery to understand and promote sustainable development. Science 371, eabe8628.

Ciccone, A., Hall, R.E., 1996. Productivity and the Density of Economic Activity. The American Economic Review 86, 54–70. URL: http://www.jstor.org/stable/2118255.

Dietterich, T., 2000. An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization. Machine Learning 40, 139–157.

Duranton, G., Puga, D., 2015. Urban Land use, in: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), Handbook of Regional and Urban Economics. Elsevier. volume 5, pp. 467–560.

Elvidge, C.D., Baugh, K., Zhizhin, M., Hsu, F.C., Ghosh, T., 2017. VIIRS night-time lights. International Journal of Remote Sensing 38, 5860–5879.

Engstrom, R., Hersh, J., Newhouse, D., 2021. Poverty from Space: Using High Resolution Satellite Imagery for Estimating Economic Well-being. The World Bank Economic Review 0, 1–31.

Friedman, J.H., 2002. Stochastic gradient boosting. Computational Statistics & Data Analysis 38, 367–378.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognition Letters 31, 2225–2236.

Glaeser, E.L., Xiong, W., 2017. Urban productivity in the developing world. Oxford Review of Economic Policy 33, 373–404.

Gobillon, L., Selod, H., 2014. Spatial Mismatch, Poverty, and Vulnerable Populations, in: Fischer, M.M., Nijkamp, P. (Eds.), Handbook of Regional Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 93–107. URL: http://dx.doi.org/10.1007/978-3-642-23430-9_7.

Goldblatt, R., Heilmann, K., Vaizman, Y., 2020. Can Medium-Resolution Satellite Imagery Measure Economic Activity at Small Geographies? Evidence from Landsat in Vietnam. The World Bank Economic Review 34, 635–653.

Gudmundsson, A., Mohajeri, N., 2013. Entropy and order in urban street networks. Scientific Reports 3, 3324.

Hallegatte, S., Rentschler, J., Rozenberg, J., 2019. Lifelines: The Resilient Infrastructure Opportunity. The World Bank. URL: http://elibrary.worldbank.org/doi/book/10.1596/978-1-4648-1430-3, doi:10.1596/978-1-4648-1430-3.

He, Y., Thies, S., Avner, P., Rentschler, J., 2021. Flood impacts on urban transit and accessibility—A case study of Kinshasa. Transportation Research Part D: Transport and Environment 96, 102889. URL: https://doi.org/10.1016/j.trd.2021.102889, doi:10.1016/j.trd.2021.102889.

Henderson, J., Regan, T., Venables, A., 2021a. Building the City: From Slums to a Modern Metropolis. The Review of Economic Studies 88, 1157–1192.

Henderson, J.V., Kriticos, S., 2018. The Development of the African System of Cities. Annual Review of Economics 10, 287–314.

Henderson, J.V., Nigmatulina, D., Kriticos, S., 2021b. Measuring urban economic density. Journal of Urban Economics 125, 103188. doi:10.1016/j.jue.2019.103188.

Henderson, J.V., Storeygard, A., Weil, D.N., 2012. Measuring Economic Growth from Outer Space. The American Economic Review 102, 994–1028. URL: http://www.jstor.org/stable/23245442.

Hengl, T., 2006. Finding the right pixel size. Computers & Geosciences 32, 1283–1298.

Hoerl, A.E., Kennard, R.W., 1970. Ridge Regression: Applications to Nonorthogonal Problems. Technometrics 12, 69.

Jarvis, A., Reuter, H., Nelson, A., Guevara, E., Guevara, A., 2008. Hole-filled SRTM for the globe Version 4.

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

Kreindler, G., Miyauchi, Y., 2021. Measuring Commuting and Economic Activity inside Cities with Cell Phone Records. Technical Report w28516. National Bureau of Economic Research. Cambridge, MA. URL: http://www.nber.org/papers/w28516.pdf, doi:10.3386/w28516.

Lall, S.V., Lebrand, M., Park, H., Sturm, D., Venables, A., 2021a. Pancakes to Pyramids : City Form to Promote Sustainable Growth. The World Bank, Washington, D.C. URL: http://documents.worldbank.org/curated/en/554671622446381555/City-Form-to-Promote-Sustainable-Growth.

Lall, S.V., Lebrand, M., Soppelsa, M.E., 2021b. The Evolution of City Form. Evidence from Satellite Data. Policy Research Working Paper 9618. The World Bank. Washington, D.C.

Li, P., Zhao, P., Schwanen, T., 2020. Effect of land use on shopping trips in station areas: Examining sensitivity to scale. Transportation Research Part A: Policy and Practice 132, 969–985.

Louail, T., Lenormand, M., Cantu Ros, O.G., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M., 2015. From mobile phone data to the spatial structure of cities. Scientific Reports 4, 5276. URL: http://www.nature.com/articles/srep05276, doi:10.1038/srep05276.

Lu, X., Bengtsson, L., Holme, P., 2012. Predictability of population displacement after the 2010 Haiti earthquake. Proceedings of the National Academy of Sciences 109, 11576–11581. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1203882109, doi:10.1073/pnas.1203882109.

Lyapustin, A., Wang, Y., 2018. MCD19A2 MODIS/Terra+Aqua Land Aerosol Optical Depth Daily L2G Global 1km SIN Grid V006; distributed by NASA EOSDIS Land Processes DAAC.

Miyauchi, Y., Nakajima, K., Redding, S., 2021. Consumption Access and Agglomeration: Evidence from Smartphone Data. Technical Report w28497. National Bureau of Economic Research. Cambridge, MA. URL: http://www.nber.org/papers/w28497.pdf, doi:10.3386/w28497.

Moran, P.A.P., 1948. The Interpretation of Statistical Maps. Biometrika 35, 255 – 60.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 2, 559–572.

Perlich, C., Provost, F., Simonoff, J., Cohen, W., 2003. Tree Induction vs. Logistic Regression: A Learning-Curve Analysis. Journal of Machine Learning Research 4, 211 – 255.

Pesaresi, M., Ehrilch, D., Florczyk, A., Freire, S., Julea, A., Kemper, T., Soille, P., Vasileios, S., 2015. GHS built-up grid, derived from Landsat, multitemporal (1975, 1990, 2000, 2014).

Pogonyi, C.G., Graham, D.J., Carbo, J.M., 2021. Metros, agglomeration and displacement. Evidence from London. Regional Science and Urban Economics 90, 103681.

Soman, S., Beukes, A., Nederhood, C., Marchio, N., Bettencourt, L., 2020. Worldwide Detection of Informal Settlements via Topological Analysis of Crowdsourced Digital Maps. ISPRS International Journal of Geo-Information 9, 685.

Straulino, D., Saldarriaga, J.C., Gómez, J.A., Duque, J.C., O'Clery, N., 2021. Uncovering commercial activity in informal cities. arXiv preprint arXiv:2104.04545 .

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society: Series B (Methodological) 58, 267–288.

U.S. Geological Survey, . Landsat 8 Collection 1 Tier 1 Annual NDWI Composite.

Varian, H.R., 2014. Big Data: New Tricks for Econometrics. Journal of Economic Perspectives 28, 3–28.

Yabe, T., Jones, N.K.W., Lozano-Gracia, N., Khan, M.F., Ukkusuri, S.V., Fraiberger, S., Montfort, A., 2021. Location Data Reveals Disproportionate Disaster Impact Amongst the Poor: A Case Study of the 2017 Puebla Earthquake Using Mobilkit. arXiv:2107.13590 [physics] URL: http://arxiv.org/abs/2107.13590. arXiv: 2107.13590.

Yabe, T., Tsubouchi, K., Fujiwara, N., Sekimoto, Y., Ukkusuri, S.V., 2020. Understanding post-disaster population recovery patterns. Journal of The Royal Society Interface 17, 20190532. URL: https://royalsocietypublishing.org/doi/10.1098/rsif.2019.0532, doi:10.1098/rsif.2019.0532.

Yeh, C., Perez, A., Driscoll, A., Azzari, G., Tang, Z., Lobell, D., Ermon, S., Burke, M., 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. Nature Communications 11, 2583.

Zagatti, G.A., Gonzalez, M., Avner, P., Lozano-Gracia, N., Brooks, C.J., Albert, M., Gray, J., Antos, S.E., Burci, P., zu Erbach-Schoenberg, E., Tatem, A.J., Wetter, E., Bengtsson, L., 2018. A trip to work: Estimation of origin and destination of commuting patterns in the main metropolitan regions of Haiti using CDR. Development Engineering 3, 133–165. URL: https://linkinghub.elsevier.com/retrieve/pii/S2352728517300866, doi:10.1016/j.deveng.2018.03.002.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67, 301–320.

Appendix A Data Source Details

Employment Data

Cities in Sub-Saharan Africa

Abidjan/Republic of Côte d'Ivoire

Japan International Cooperation Agency (JICA), Ministry of Construction, Housing, Sanitation and Urban Development of the Republic of Côte d'Ivoire (MCLAU) and Schema Directeur d'Urbanisme du Grand Abidjan (2015). The Project for the Development of the Urban Master Plan in Greater Abidjan (SDUGA).

Dakar/Republic of Senegal

The World Bank Group, Conseil Executif des Transports Urbains de Dakar (CETUD) and Consortium Solidarité Internationale Sur les Transports et la Recherche en Afrique Sub-Saharienne (2015). Enquête Mobilité, Transports et Accès aux Services Urbains de Dakar (EMTASUD).

Dar es Salaam/United Republic of Tanzania

Not yet known

Douala/Republic of Cameroon

Agence Française de Développement (AFD), European Commission, Syndicat Mixte des Transports pour le Rhône et l'Agglomération (SYTRAL) and MobiliseYourCity Partnership (2018). Enquête Ménage-Déplacements as part of the Plan de Mobilité Urbaine Soutenable (PMUS)/Communaute Urbaine de Douala.

Harare/Republic of Zimbabwe

Zimbabwe National Statistics Agency (ZimStat), United Nations Population Fund (UNFPA), UK Department for International Development (DFID), Australian Agency for International Development, Danish International Development Agency, United Nations Children's Fund (UNICEF), European Union, The Swedish International Development Cooperation Agency (SIDA) and United Nations Development Program (UNDP) (2012). Zimbabwe Population Census 2012.

Kampala/Republic of Uganda

Travel Survey
Kampala Capital City Authority (KCCA), ROM Transportation Engineering Ltd., Shapira-Hellerman Planners, Larry Aberman & Associates, Tzamir Architects and Planners Ltd. and Ofek Aerial Photography (2012). Kampala Physical Development Plan (KPDP).
Firm Census
Uganda Bureau of Statistics (UBOS) (2011). Census of Business Establishments (COBE).

Kigali/Republic of Rwanda

National Institute of Statistics of Rwanda (NISR) (2011). Rwanda Establishment Census.

Kinshasa/Democratic Republic of Congo

Japan International Cooperation Agency (JICA), ALMEC Corporation and Oriental Consultants Global Co., Ltd (2018). Kinshasa Commuter Travel Survey (CTS).

Nairobi/Republic of Kenya

Japan International Cooperation Agency (JICA), Nairobi City Council (NCC), Nippon Koei Co., Ltd., IDCJ Inc. and EJEC Inc. (2013). The Project on Integrated Urban Devel-

opment Master Plan for the City of Nairobi in the Republic of Kenya - Nairobi Personal Travel Survey.

Cities in Latin America

Belo Horizonte/Federative Republic of Brazil

Brazilian Ministry of the Economy (2019). Relação Anual de Informações Sociais (RAIS).

Bogotá/Republic of Colombia

Not yet known

(Autonomous City of) Buenos Aires/Argentine Republic

Secretariat of Transport of the Argentinian Ministry for Federal Planning, Public Investments, and Services (MINPLAN), Ingeniería en Relevamientos Viales S.A. (IRV), Ingeniería y Asistencia Técnica Argentina S.A. (IATSA) and LOGIT - Transportation Engineers (2012). 2004/5 Censo Nacional Ecónomico/The Argentinian National Institute of Statistics and Census (INDEC) and 2011 Encuesta Permanente de Hogares/INDEC.

Lima/Republic of Peru

Peruvian National Institute of Statistics and Informatics (INEI) (2010). Census of Population and Housing Units.

Mexico City/United Mexican States

Mexican National Statistical Directory of Economic Units (DENUE) (2019). The Business Register of Mexico (RENEM).

Satellite Data

Night Lights

VIIRS Stray Light Corrected Nighttime Day/Night Band Composites Version 1; 15 arc seconds resolution (Elvidge et al., 2017)

Population

WorldPop Global Project Population Data: Estimated Residential Population per 100 x 100m Grid Square; 100m^2 resolution

Air Pollution

Terra + Aqua MAIAC Land Aerosol Optical Depth Daily; 1km^2 resolution (Lyapustin and Wang, 2018)

Geophysical Land Surface

SRTM Digital Elevation Data Version 4; 90m^2 resolution (Jarvis et al., 2008)

Built-up Land Cover

GHSL: Global Human Settlement Layers, Built-Up Grid 1975-1990-2000-2015 (P2016); 38m^2 resolution (Pesaresi et al., 2015)

Normalized Difference Vegetation Index/NDVI

Landsat 8 Collection 1 Tier 1 32-Day NDVI Composite; 30m^2 resolution

Normalized Difference Water Index/NDWI (U.S. Geological Survey)

Landsat 8 Collection 1 Tier 1 Annual NDWI Composite; 30m^2 resolution (U.S. Geological Survey)

Land Use

Copernicus Global Land Cover Layers: CGLS-LC100 Collection 3; 100m^2 resolution (Buchhorn et al., 2020)

Appendix B Additional Tests

This section provides additional computations to investigate possible effects of data collection and of choice of different spatial units.

B.1 Heterogeneity in Employment Data Collection

The employment data used throughout this work is collected either via travel surveys or via population and firm censuses[16]. This introduces noise to the data both with respect to heterogeneity of polygon sizes within which the data was originally collected but also in terms of what specific employment is measured. Travel surveys overall typically include any type of employment, whereas firm censuses are generally limited to collecting information on formal employment and do not capture employment within the informal sector. As the informal sector represents a substantial part of jobs within developing countries (Bryan et al., 2020), omitting this might introduce a noticeable measurement issue within the training of the algorithms. In order to test this, we split the cities according to employment data collection type, as outlined in Table 1, and retrain the Spatial RF algorithm for each of the two types of employment data collection. Results are provided in Table 6.

Across both types of employment data collection, the predictive performance is relatively similar with slightly better predictive performance for employment data derived via

[16] See Table 1.

surveys. When evaluating the models' performance via the data obtained through the alternative data collection method, the model trained on census data appears to predict employment density collected via survey data slightly better than census data. This might indicate that the measurement error introduced through neglecting informal employment within census data weakens the relationship between the dependent variable and the features and thus results in lower predictive performance. However, given the relatively high predictive performance for each type of data collection and the relatively low predictive performance when evaluated on data obtained via the alternative, the results are more likely point to the importance of similarity of data used for the train and test sets. Furthermore, given that the employment data for each city is collected either through a census or through a survey, we cannot exclude that the results are driven by city idiosyncrasy.

In order to test this further, we evaluate both models specifically for data for Kampala/UGA where we have access to separate employment data obtained via both options. The evaluation of the trained models reveals a predictive performance of R^2 of 0.87 and 0.68 for survey and census data respectively with an RMSE of 0.38 for the former and 0.57 for the latter. This further strengthens the notion that survey data might be a more suitable data source for employment prediction.

However, given that observations of each of the Kampala data was included within the training data for each model, we retrain the equivalent algorithms on survey and census data cities fully excluding grid cells located in Kampala. While both the model based on survey and on census data lose substantial predictive performance, the model trained on census data outperforms that trained on survey data with R^2 of 0.53 and 0.39 respectively. Hence, while survey data appears to explain and predict employment data better when there exists no structural differences across the train and test data set, census data appears to suffer from a lower out-of-sample bias when evaluated on data that differs structurally from the train data used for algorithms identification.

B.2 Sensitivity of Results to Levels Spatial (Dis)Aggregation

The choice of spatial boundaries generally impacts the results of quantitative spatial studies (Hengl, 2006; Li et al., 2020) driven by both the exact placement of the spatial unit's border and/or spatial (dis)aggregation within the spatial unit's value. While the possible bias introduced through both, issues might already be partly lessened through our incorporation of neighboring grid cells' values. Here we explicitly test the sensitivity of our trained algorithms by extracting grid cells values of the dependent variable and all features with alternative hexagonal grid dimensions: 250m x 250m, 1km x 1km and 2km x 2km[17]. We additionally retrain the algorithm using the original polygon boundaries as observations. Table 7 provides

[17] A hexagonal grid cell structure of 1.5km x 1.5km dimension has also been tested but results are omitted here.

Table 6: Performance Comparison across RF Models with Spatial Effects (R^2 (RMSE)) Data Collection Subsamples

		Train Data	
		Survey Data	Census Data
Test Data	Survey Data	0.88 (0.35)	0.47 (0.73)
	Census Data	0.40 (0.80)	0.83 (0.42)
No. Obs.		75,543	19,336

Note: Cross-Evaluation has been conducted on the Complete Survey and Census Data; No. Obs. refers to the full data sets of both train and test data for each sample;

the results.

Comparing the performance of the algorithms as evaluated on the structurally identical test and train data across the three grid cell dimensions shows that the algorithm based on 250m x 250m grid cells delivers the best predictive performance across all observations and those based on geographical region. However, this remains below the performance quality of the algorithm trained on 500m x 500m grid cells. In contrast, cross-evaluation points to better predictions when larger grid cells are used, although this benefit appears to become marginally smaller with grid cells structures larger than 1km x 1km. As discussed throughout Section 5, algorithms are biased towards the mean and predict outliers less accurate. When we use larger grid cells, the underlying data is spatially averaged to compute grid cell values which results in less extreme outlier values when larger grid cells are used. This is likely contributing to better predictive performance for larger grid cell units when we evaluate it on a test data set that is structurally different from the train data used to fit the model. This is in line with the previous results indicating that better predictions on the structurally identical test data are generally associated with lower predictive performance on data that exhibits structural differences. Further, these results do not point to the introduction of a spatial disaggregation bias which would add statistical noise to smaller grid cell dimensions and subsequently lead to reductions in predictive performance. However, it cannot be excluded that the higher predictive performance for smaller units is influenced by the larger number of observations underlying the training of the algorithm.

The algorithm trained on the polygon data overall delivers relatively strong predictive performance for observations of Latin America; however, it yields noticeably lower predictive performance across SSA. The employment data for LAT cities has been provided in on average smaller polygons, resulting in a substantially higher amount of observations and thus benefiting the accuracy of the algorithm. Despite this, the quality of the predictions identified

Table 7: Performance Comparison across RF Models with Spatial Effects (R^2 (RMSE))
Different levels of spatial disaggregation

		Train Data					
		All Cities	SSA Cities	LAT Cities	All Cities	SSA Cities	LAT Cities
		Raw Polygons			250m x 250m Grid Cells		
Test Data	All Cities	0.79 (0.46)	0.40 (0.81)	0.71 (0.55)	0.88 (0.35)	0.59 (0.67)	0.41 (0.79)
	SSA Cities	0.61 (0.62)	0.54 (0.70)	0.47 (0.78)	0.90 (0.33)	0.91 (0.31)	0.36 (0.83)
	LAT Cities	0.80 (0.45)	0.39 (0.80)	0.81 (0.44)	0.86 (0.38)	0.51 (0.73)	0.86 (0.38)
No. Obs.		38,090	2,733	35,357	385,458	221,890	163,568
		1km x 1km Grid Cells			2km x 2km Grid Cells		
Test Data	All Cities	0.83 (0.42)	0.63 (0.63)	0.50 (0.76)	0.76 (0.47)	0.63 (0.59)	0.51 (0.73)
	SSA Cities	0.84 (0.42)	0.82 (0.44)	0.43 (0.82)	0.74 (0.50)	0.76 (0.47)	0.46 (0.77)
	LAT Cities	0.83 (0.42)	0.59 (0.68)	0.82 (0.43)	0.79 (0.43)	0.55 (0.67)	0.77 (0.46)
No. Obs.		25,578	14,959	10,619	6,832	4,092	2,740

Note: Train and test data sets have been obtained via a random 80/20 split; No. Obs. refers to the full data sets of both train and test data for each sample

for LAT remains below that identified for the majority of grid cell dimensions.

Overall, these results do not provide support for the hypothesis that spatial (dis)aggregation has introduced a substantial bias into the computations based on 500m x 500m grid cells. Furthermore, smaller units lead to better predictions in cases where test and train data are structurally identical. However, this might be strongly influenced by the larger training data set.

Appendix C Additional Visuals

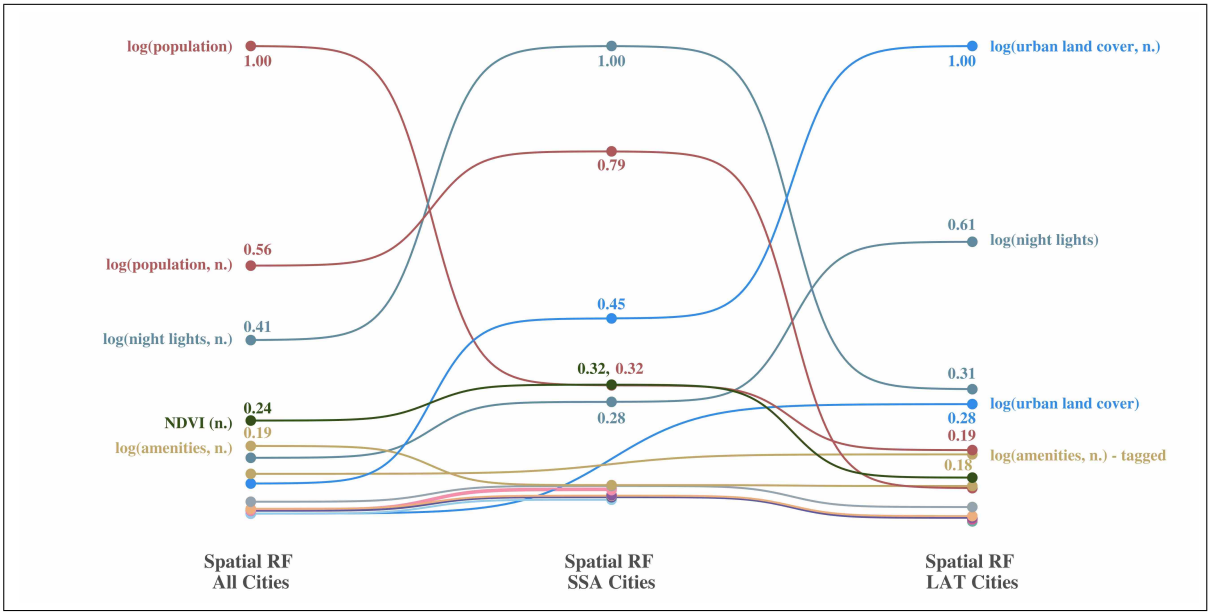

Figure 12: Relative Variable Importance across Spatial RF models

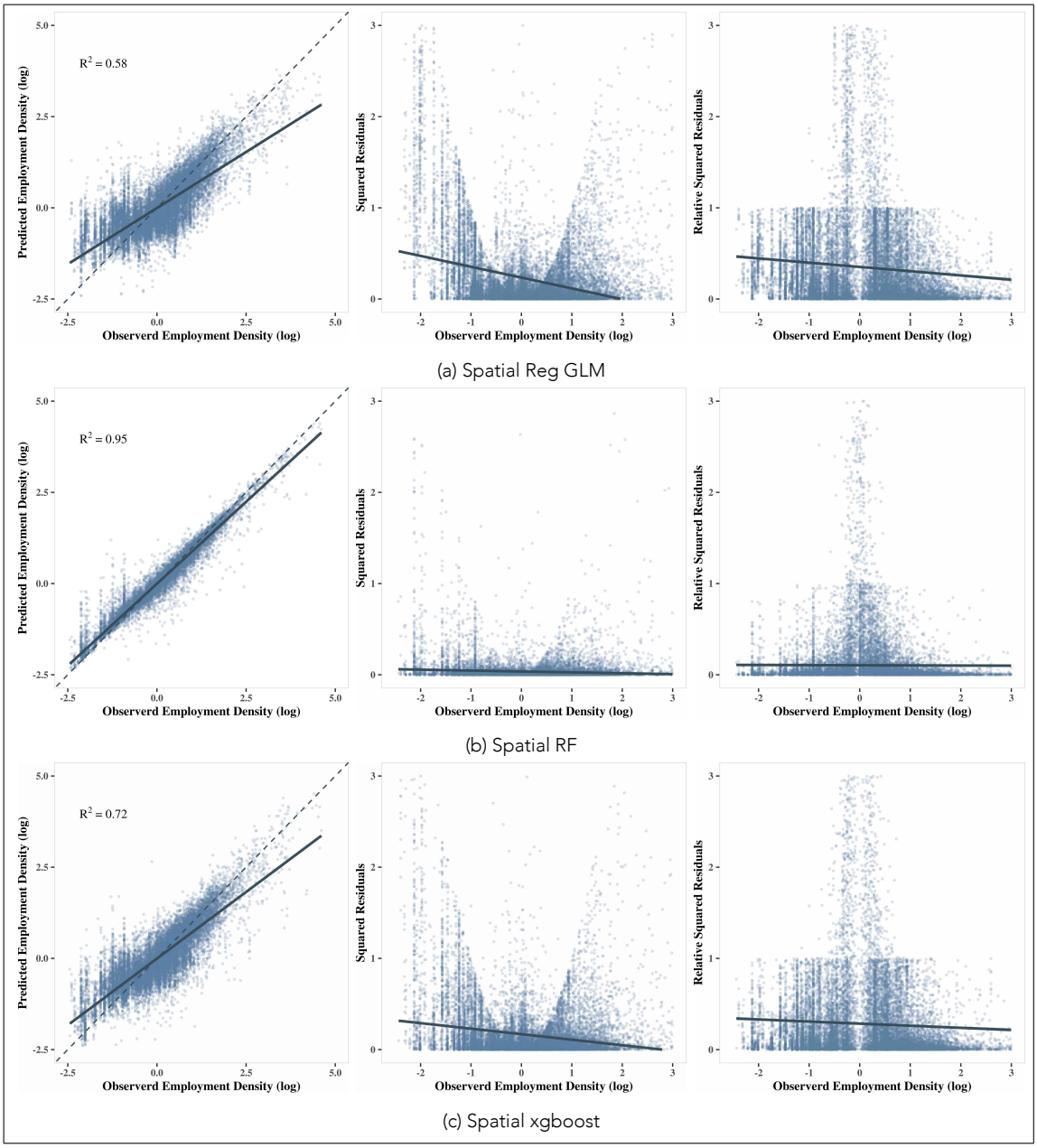

Figure 13: Illustration of Performance Metrics across Algorithms