PREDICTING URBAN EMPLOYMENT DISTRIBUTIONS

A toolkit for more targeted urban investment and planning decisions

The challenge: employment density maps, key to targeting urban investments, are often outdated, imprecise, or unavailable

Information on the spatial distribution of jobs, both formal and informal, in urban areas is a fundamental requirement for many project appraisals and analyses. Such information allows urban planners and developers to identify economic hubs within a city and take targeted measures to improve their productivity, connectivity, and resilience—for example, by investing in infrastructure upgrades or flood protection systems, enhancing commuting options, and adapting urban planning decisions. Such measures support firms and yield city-wide benefits for the lives and livelihoods of workers and their communities.

In practice, business registries, employment censuses, or travel surveys are the most common sources for mapping the density and spatial distribution of jobs within a city. But such data are rarely available; and when they do exist, tend to be incomplete, unreliable, or outdated. Recent initiatives have successfully leveraged mobile phone-derived data to document "meaningful" locations, including jobs. This is a breakthrough, especially given the increasingly ubiquitous use of mobile phones. And yet, accessing and processing mobile phone data is a difficult, lengthy, and often costly process.

In this note, we demonstrate a machine learning approach for high-resolution urban employment prediction using a robust and scalable approximation methodology that relies on widely available public data, making it a quick, low-cost solution. We show that in most developing countries, this approach can open new avenues for targeted urban investment and planning decisions that are based on systematic empirical evidence.

An analytical solution, in a nutshell

A machine learning approach for high-resolution urban employment prediction in developing countries

Relying on open-source data extracted from OpenStreetMap (OSM) and Google Earth Engine (GEE), we provide a new analytical toolkit to approximate for the spatial distribution of jobs in urban developing country areas. Using machine learning algorithms, we show that it is possible to predict employment density based on urban form attributes, such as street density and amenities. Using this approach, we generate predicted employment density maps for 14 cities in Latin America and Sub-Saharan Africa, from Dakar in Senegal to Buenos Aires in Argentina (figure 1), validating the robustness of these maps against survey-based observed employment density data. Generally, we find that the approach predicts within-city employment concentrations with high resolution and accuracy, and can therefore be replicated in cities with no observed employment information to inform urban investments and planning decisions.

Figure 1. Observed and predicted employment density in urban Buenos Aires, Argentina (R^2=0.78)

(a) Observed employment density

(b) Predicted employment density

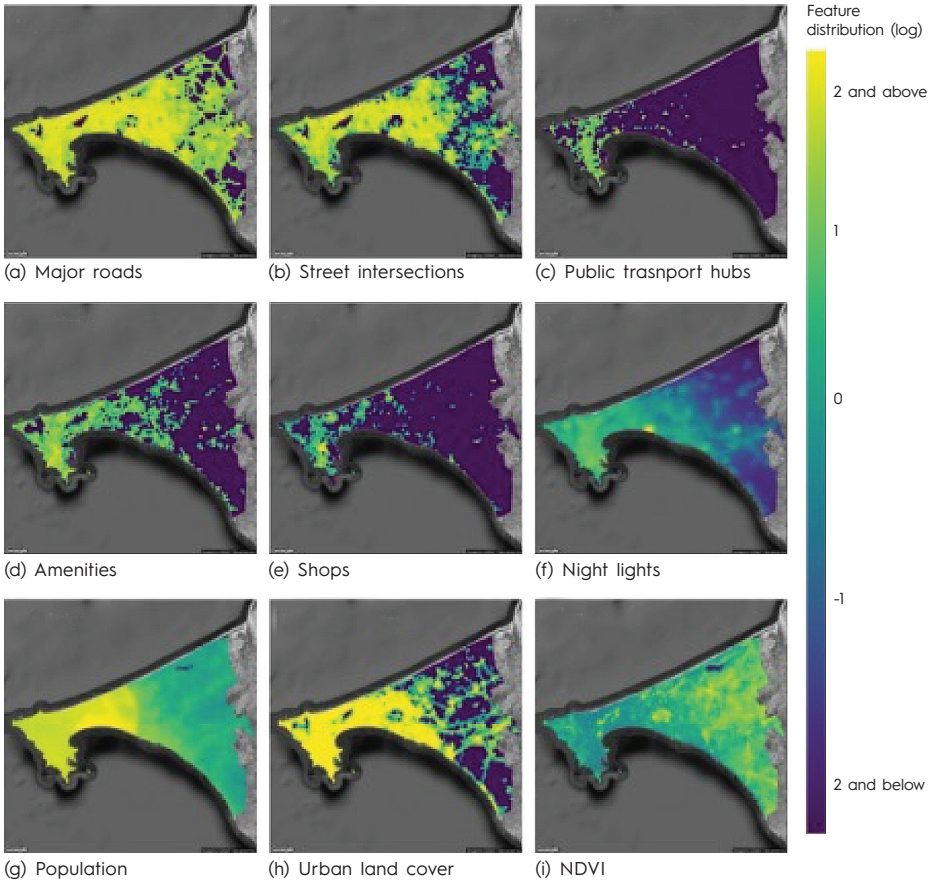
Note: R^2, or goodness of fit measure, indicates the share of variation in the observed employment density that can be explained by the algorithm.

GFDRR Global Facility for Disaster Reduction and Recovery
Administered by THE WORLD BANK IBRD • IDA | WORLD BANK GROUP

Methodology

We exploit existing correlations between urban form properties—such as nighttime light brightness and road intersection density—and job density to predict the spatial distribution of employment in urban areas of developing country cities. Our methodology relies entirely on open-source data leveraged from OSM and GEE and is tested for 14 cities in Latin America and Sub-Saharan Africa, chosen for having available employment data through business registries or travel surveys, which enabled us to validate predictions.

Figure 2. Illustration of selected features at 500x500-meter hexagonal grid cells



Note: Panels (a) to (e) are density-adjusted; Normalized Difference Vegetation Index (NDVI) is not logarithm-adjusted; all features are within-city scaled.

First, we use several methodologies to understand the explanatory power of various features of the data derived from OSM and GEE to predict employment distribution in the 14 urban areas. We find that population, nighttime lights, urban land cover, amenities, and road intersections are strongly and positively correlated to job locations, and that terrain roughness, water bodies, and vegetation indices are strongly and negatively correlated to the presence of jobs.

Second, we test several model specifications for employment predictions at grid cell level, including Penalized Linear Regression models and Ensemble Tree methods. We settle on a spatial variation of the Random Forest machine learning algorithm that can capture the clustering of economic activity to predict employment density at grid cell level.

Applications and insights

The approach enables two types of predictive analysis. The first predicts employment distribution in cities for which we have no employment data. The second predicts employment in cities for which we have some employment data and incomplete spatial coverage.

When there is no employment data

The first type of analysis aims to understand how well our methodology can predict employment density at grid cell level for entirely "unseen" cities, where the algorithm has not been trained and we have no employment information. To do this, we train the spatial random forest algorithm on data from 13 cities and hold back data from the city we are trying to predict. We repeat this 14 times, each time holding back data from a different city, to predict employment density for all the cities in the study. The R^2—that is, the goodness of fit measure—indicates the share of variation in the observed employment density that can be explained by the predictions derived from the algorithm (Table 1).

We find that our method can predict employment density for polygons in out-of-sample cities with medium to high accuracy (mean R^2 is 0.63, and maximum R^2 is 0.81). However, the results show heterogeneity in predictive performance, with R^2s ranging

Figure 2. Performance comparison across Random Forest models with spatial effects for out-of-sample cities (R^2)

Sub-Saharan Africa								
Abidjan	Dakar	Dar Es Salaam	Douala	Harare	Kampala	Kigali	Kinshasa	Nairobi
0.70	0.70	0.71	0.65	0.30	0.54	0.81	0.55	0.77

Latin America				
Belo Horizonte	Bogota	Buenos Aires	Lima	Mexico City
0.77	0.52	0.78	0.42	0.58

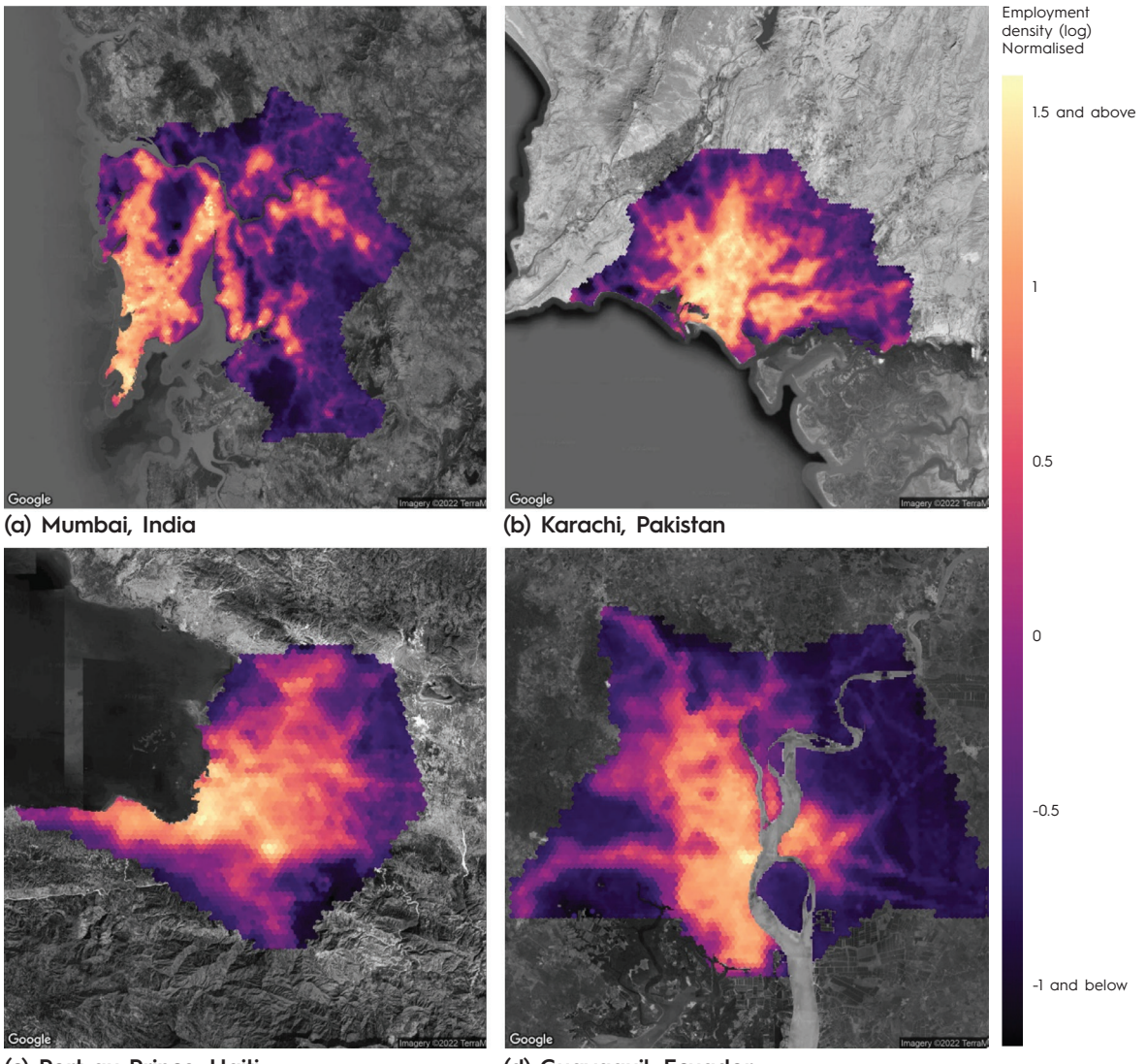
from 0.30 for Harare to 0.81 for Kigali. Comparing the results by geographical region reveals that Latin American and Sub Saharan African cities are equally well predicted on average, with mean R^2 of 0.61 and 0.64 respectively. These results are in line with, but above, a similar analysis using satellite imagery undertaken to predict consumption expenditure and household assets of spatial clusters across four SSA countries[1] and substantially above those obtained in the closest related literature estimating employment factors which obtained R^2s ranging from 0.24 to 0.33.[2]

Are the city level R^2 differences a reason for concern? We can attribute the variation in out-of-sample prediction to a combination of city-specific relationships between employment and OSM and GEE features, and varying degrees of quality for employment and feature data across cities.

[1] Jean, Neal, Marshall Burke, Michael Xie, W. Matthew Davis, David B. Lobell, and Stefano Ermon. 2016. "Combining Satellite Imagery and Machine Learning to Predict Poverty." *Science* 353 (6301): 790–94. https://doi.org/10.1126/science.aaf7894.

[2] Goldblatt, Ran, Kilian Heilmann, and Yonatan Vaizman. 2020. "Can Medium-Resolution Satellite Imagery Measure Economic Activity at Small Geographies? Evidence from Landsat in Vietnam." *The World Bank Economic Review* 34 (3): 635–53. https://doi.org/10.1093/wber/lhz001.

Figure 3. Empoyment predictions across various cities

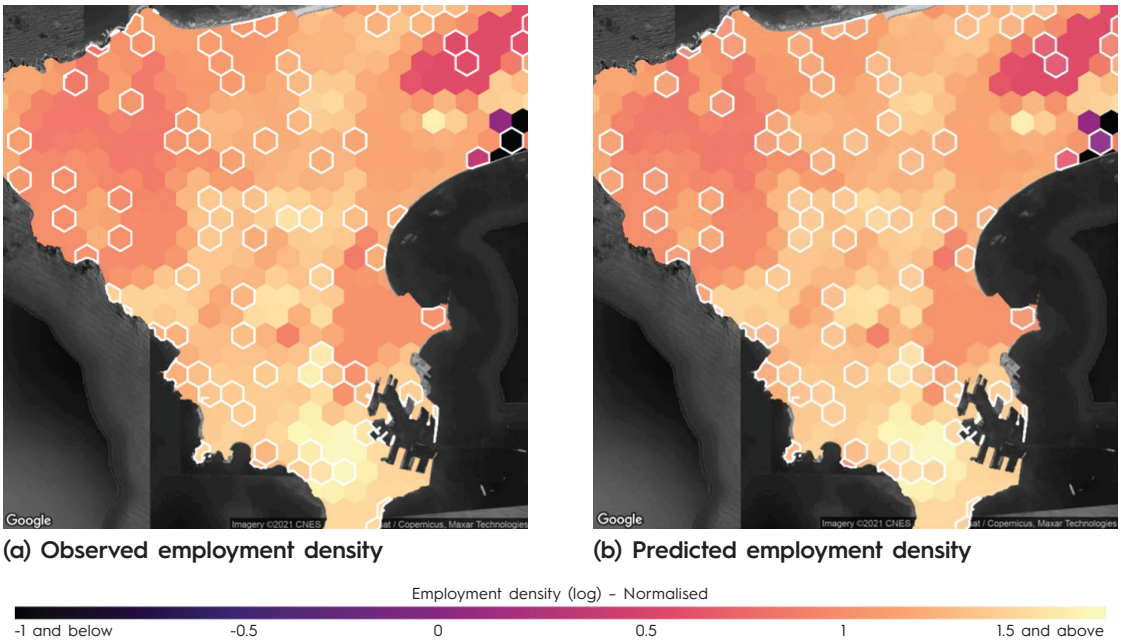


(a) Mumbai, India (b) Karachi, Pakistan

(c) Port au Prince, Haiti (d) Guayaquil, Ecuador

Figure 4. Observed and predicted values for test data grid cells in urban Dakar, Senegal, using spatial Random Forest models

Note: Test cells are outlined in white



(a) Observed employment density

(b) Predicted employment density

When information is patchy

The second type of analysis trains the spatial Random Forest algorithm on 80 percent of all grid cells, pooling grid cells from the 14 cities for which we have employment data. It then applies the trained algorithm on the 20 percent remaining grid cells in each city. This analysis is useful when some information on employment is available at the city level but spatially incomplete. The algorithm is therefore trained on existing city level data. Our algorithm's predictive performance is measured by the quality of fit between the predicted and observed value of employment density in the 20 percent "untrained" grid cells. We find that our method can predict employment density in held-back within-city cells in our test cities with extremely high accuracy, as measured by an R^2 higher than 0.95 (figure 4). This shows that when we have spatially "patchy" or incomplete city-level employment data, we can approximate for employment density in the areas without data.

Possible applications of the methodology

Our methodology's satisfactory performance in predicting spatial employment distribution in "data patchy" or "unseen" cities opens up possibilities for analytical and operational applications. First, it allows for more systematic employment accessibility analyses in developing country cities. Such analyses aim to better measure the benefits of, and tailor, transport investments and land use interventions, and are mandatory for World Bank urban transport interventions. Second, it could help increase the development and application of quantitative spatial economic models, which need location and volume of employment as inputs. And third, it could help document the spatial structure of urban areas and improve the measurement of agglomeration forces in developing country cities.

Future work and caveats

We have tested this version of our methodology on Sub-Saharan and Latin American cities. Future efforts should replicate this work in the other World Bank regions: East Asia and Pacific, Europe and Central Asia, Middle East and North Africa, and South Asia. Additional time and efforts should also be invested in better understanding the heterogenous predictive performance of the machine learning algorithms for "unseen" cities. When alternative data, such as employment data derived from employment censuses, travel surveys, or even mobile phones, is available, it should be preferred. We equally caution against the use of this prediction methodology to measure the evolution of job distribution over time without further validation.

Further reading

The interested reader is welcome to explore our technical paper:

Barzin, Samira, Paolo Avner, Jun Rentschler, and Neave O'Clery. 2022. "Where Are All the Jobs ?: A Machine Learning Approach for High Resolution Urban Employment Prediction in Developing Countries." World Bank Policy Research Working Paper, no. 9979. https://openknowledge.worldbank.org/handle/10986/37195

Contacts

The Global Facility for Disaster Reduction and Recovery's (GFDRR) Global Programs on Resilient Infrastructure and Disaster Risk Analytics can provide support in applying the operational analytics approach presented in this note to urban infrastructure and resilience projects.

For more information, or if you are interested in applying this methodology to your projects or analyses, please feel free to contact:

- **Paolo Avner,** Urban Economist, GFDRR: pavner@worldbank.org
- **Jun Rentschler,** Senior Economist, Office of the Chief Economist for Sustainable Development: jrentschler@worldbank.org