Occupations on the map: Using a super learner algorithm to downscale labor statistics

Michiel van Dijk (✉ michiel.vandijk@wur.nl)

Wageningen Economic Research <https://orcid.org/0000-0002-5207-7304>

Thijs de Lange

Wageningen Economic Research <https://orcid.org/0000-0003-1599-7037>

Paul van Leeuwen

Wageningen Economic Research <https://orcid.org/0000-0002-9784-8037>

Philippe Debie

Wageningen Economic Research <https://orcid.org/0000-0002-7353-9715>

Research Article

Keywords: machine learning, labor, super learner, downscaling, occupation

Posted Date: January 31st, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1300541/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Occupations on the map: Using a super learner algorithm to downscale labor statistics

Michiel van Dijk^{1,2,*} Thijs de Lange¹ Paul van Leeuwen¹ Philippe Debie^{1,3}

2022-01-26

¹ Wageningen Economic Research, the Hague, the Netherlands

² International Institute for Applied Systems Analysis

³ Marketing and Consumer Behaviour Group, Wageningen University, Wageningen, the Netherlands

* Correspondence: Michiel van Dijk <michiel.van.dijk@wur.nl>

Abstract

Detailed and accurate labor market statistics are fundamental to support social policies that aim to improve the match between labor supply and demand, and support the creation of jobs. Despite overwhelming evidence that labor activities are distributed unevenly across space, detailed statistics on the geographical distribution of labor and work are not readily available. To fill this gap, we demonstrated an approach to create fine-scale gridded occupation maps by means of downscaling district-level labor statistics informed by remote sensing and other spatial information. We applied a super learner algorithm that combines the results of different machine learning models to predict the shares of six major occupation categories and the labor force participation rate at a resolution of 30 arc seconds (1x1 km) in Vietnam. The results are subsequently combined with gridded information on the working-age population to produce maps of the number of workers per occupation. The super learners outperform ($n = 4$) or have similar ($n = 3$) accuracy in comparison to best-performing single machine learning algorithms. A comparison with an independent high-resolution wealth index showed that the shares of the four low-skilled occupation categories (91% of the labor force), were able to explain between 27% and 45% of the spatial variation in wealth in Vietnam, pointing at a strong spatial relationship between work, income and wealth. The proposed downscaling approach can also be applied to produce maps of other (labor) statistics, which are only available at aggregated levels.

1 Introduction

Labor is recognized as one of the three primary factors of production in economics and accounts for around 50% of total global income (Karabarbounis et al., 2014). Work forms a central part of most people’s lives. Globally more than 3.1 billion people are working actively or looking for work (ILO, 2021) and those that are employed spend around 40 hours per week on the job (Giattino et al., 2013). There is overwhelming evidence that labor activities are distributed unevenly across space because of a combination of differences in costs, economies of scale and spillover effects (Ellison et al., 1999; Krugman, 1991; Rickard, 2020). However, detailed statistics that illustrate the geographical distribution of labor and work are not readily available. ILOSTAT (ilostat.ilo.org), the most comprehensive global labor statistics database maintained by the International Labor Office (ILO) only provides country-level data. Summary reports of population censuses and national labor-force surveys sometimes present subnational labor information, but mostly at the level of coarse first-level administrative units. The availability of spatially explicit labor statistics will support the formulation of targeted social policies that aim to improve the match between local labor supply and demand, and support local employment, contributing to economic growth and welfare.

The aim of this paper is to fill this gap by demonstrating an approach to create fine-scale gridded occupation maps by means of ‘downscaling’ district-level labor information using a super learner algorithm in combination with high-resolution remote sensing data and other spatial predictors (Figure 1). A super learner is an

ensemble machine learning approach that combines the results of different machine learning models to generate predictions with the same or higher accuracy than those of single models (Laan et al., 2007). Apart from better performance, combining the outcomes of multiple machine learning models also results in more robust outcomes (Polley et al., 2010), which is particularly important in case machine learning techniques are used to extrapolate results (Hengl, 2021), such as in our application.

To estimate the number of workers per occupation category, we broke it down into three components: (a) the labor force participation rate, (b) the share of workers per occupation category and (c) the working-age population, which all three have a distinct spatial distribution (Figure S2). Labor force participation rates are higher in remote areas and lower in the major cities, such as Hanoi and Ho Chi Minh City, and the more densely populated coastal areas. This is consistent with the patterns in other countries, which show that labor force participation, in particular for women, is higher in rural areas because of limited possibilities to participate in educational activities and more possibilities to combine child-rearing and farm work in the absence of off-farm employment (Mammen et al., 2000). Similarly, the share of non-agricultural occupations and the size of the working-age population tend to be higher in densely populated urban areas.

We selected Vietnam as a case study to demonstrate our approach. Vietnam is a lower middle-income country that has experienced a phase of rapid economic growth and structural transformation (Tarp, 2017). As a consequence, the occupational distribution is rather diverse, including roughly equal shares of agricultural and non-agricultural workers as well as a fair share of high-skilled jobs (Figure S1). Another advantage was the availability of relatively detailed subnational occupation statistics, which were an essential input for the analysis.

The proposed approach provides insights into the geographical distribution of workers within a country. The resulting maps make it possible to identify regions that are characterized by low-skilled employment, which are vulnerable to rising imports from low-cost countries (Autor et al., 2013). Similarly, they can be combined with climate change projections to show the locations where workers will be most exposed to extreme temperatures, which is particularly relevant for active workers in the agriculture, construction, and manufacturing sectors (Graff Zivin et al., 2014). This type of information can be used to better assess the impact of climate change on labor productivity and associated losses in national income (Lima et al., 2021; Orlov et al., 2020). The occupation maps can also be regarded as proxies for the geographical distribution of income in a country as there is a strong correlation between occupational attainment and wages (Gibbons et al., 2005). Finally, the maps also provide broad information on the spatial distribution of industrial activity as workers tend to live close to the workplace (O’Kelly et al., 2005; Sang et al., 2011), and the occupation categories are closely related to the main economic sectors: agriculture, manufacturing and services (Duernecker et al., 2021).

This paper contributes to a rapidly expanding literature, which uses machine learning and advanced statistical models to create fine-scale gridded maps of socio-economic indicators. Key examples include the mapping of population (Leyk et al., 2019), educational attainment (Graetz et al., 2019), child growth (Osgood-Zimmerman et al., 2018), poverty (Pokhriyal et al., 2017) and wealth (Chi et al., 2022). Within this field, our approach resembles the work of Stevens et al. (2015) and Nicolas et al. (2016), who disaggregated subnational information to create gridded population and livestock maps but only applied a single type (random forest) of machine learning model.

2 Results

2.1 Number of workers per occupation

Figure 2 presents maps of the number of workers for each of the six occupation categories for the entire country, and around Hanoi and Ho Chi Minh City, the two regions with the highest (working age) population density (Figure S2g). The maps clearly show different spatial patterns between the occupations. A relatively large number of agricultural workers are located in the Red River (South-East of Hanoi) and Mekong (South-West of Ho Chi Minh City) River Deltas as well as in the rural areas throughout the country. The high absolute number of agricultural workers very close to Hanoi can be explained by the combination of a large working-age population, widespread rice cultivation and well-developed urban agriculture (Lee et al., 2010). As expected, the share of agricultural workers is very small (0-2%) in urban areas and large in

Figure 1: Overview of analytical approach. (a) Subnational statistics on the labor force participation rate and occupation shares for (b) six occupation classes and (c) a large number of geospatial predictors were combined to train (d) super learner models. (e) Spatial information on the working-age population was combined with (f) resulting gridded predictions to produce (g) maps of the number of workers per occupation category. This figure has been designed using resources from Flaticon.com

rural areas (50-100%) (Figure S10). The spatial distribution of the other five occupation groups is mainly concentrated in urban and semi-urban areas. Large numbers of clerks and service workers, and craft workers and operators are depicted in the centers of the large cities but also in the industrial and populated areas that are spread around Hanoi and Ho Chi Minh City, as well as the urban coastal regions. This finding is to be expected as these two occupation categories cover the majority of jobs in the manufacturing and service sectors that employ most workers in industrial and populated areas in Vietnam. In contrast, the distribution of managers and professionals, and technicians and associate professionals is much more concentrated, with a very large presence in the centers of Hanoi and Ho Chi Minh City. These findings can be explained by the fact that high-skilled jobs are mainly concentrated in the head offices of large companies and ministries that are located in the center of major cities. In relative terms, the two high-skilled occupation categories are observed in all (semi)urban areas but only make up very small shares of the total labor force (Figure S10). Elementary occupations contain a diverse group of jobs, including both unskilled agricultural and non-agricultural workers and therefore might be allocated in rural and urban areas. The maps show that, at least in Vietnam, elementary occupations mostly seem to be of the urban type (e.g. cleaners, hand packers and street vendors).

Figure 2: Spatial predictions for the number of (a) Managers and professionals, (b) Technicians and associate professionals, (c) Clerks and service workers, (d) Agricultural workers, (e) Craft workers and operators and (f) Elementary occupations per grid cell. Bottom panels depict the spatial distribution of the number of workers around Hanoi and Ho Chi Minh City, respectively.

2.2 Model evaluation

Figure 3 summarizes the performance of the seven super learner models that were run for each variable. The amount of variation (R^2) explained by the super learner model ranged from 0.71 for elementary occupations to 0.93 for agricultural workers. A comparison of the root mean squared error (RMSE) and coefficient of determination (R^2) between the super learners and selected member models showed that the ensemble approach outperformed ($n = 4$), or was comparable to ($n = 3$) the predictions of the individual models with the highest RMSE. Out of the seven super learner models that were fitted, xgboost member models ($n = 5$), random forest ($n = 1$) and polynomial support vector machines ($n = 1$) showed the best performance (Table

S3-S9). Predictions are located around the 1:1 line and the mean error is close to zero, indicating that the models are not biased.

We also investigated if the models are able to adequately reproduce the number of workers at higher subnational aggregations. To do this, for each occupation category, we aggregated the number of workers at grid cell level and compared this with the district-level number of workers that can be derived from the model input data (Figure S2). The high R^2 of 0.89-0.96 (Figure S12) indicates a close correspondence between model outcomes and observed subnational statistics.

To provide an indication of where the super learners are accurate and where they are not, we calculated the prediction error at grid cell level (Figure S11). The maps show that the errors are most prominent in the regions where the occupation shares are very small, or even zero (i.e. agricultural workers in urban areas and non-agricultural workers in rural areas) and the models were used to extrapolate on the edges or even outside the training space.

Figure 3: Performance of the super learners for the six target variables: (a-c) Managers and professionals, (d-f) Technicians and associate professionals, (g-i) Clerks and service workers, (j-l) Agricultural workers, (m-o) Craft workers and operators, (p-r) elementary occupations and (s-u) labor force participation rate. The large panels compare the district-level observations with the super learner predictions for the hold-out dataset. The dashed blue line represents the 1:1 line. Small panels show the RMSE and the R^2 for the super learners (red lines) and super learner member models (gray lines) for the hold-out dataset. Predictions and observations are logit transformed values.

To investigate the importance of the predictor variables on the model outcomes, we conducted a variable importance analysis to investigate which predictors contributed the most to the model outcomes (Figure S13-S19) and we generated accumulated local effects plots to get insight on how they were related (Figure S20-26). Urbanization was a key predictor of occupation share as related predictors (urbp_x_prp_5 and esaccilc_dst190) were consistently ranked among the most important variables that explain model results. Another important and related determinant was night lights (viirs and dmisp). Several predictors related to land use and economic activity frequently featured as explanatory variables in some of the occupation models. Not surprisingly, for agricultural workers, the distance to cropland (esaccilc_dst011), elevation (srtm_topo) and climate (bio_5, bio_6 and bio_12) were also important determinants. In the models for

managers and professionals, and technicians and associate professionals, distance to (international) airports and travel time were frequently included, which might indicate that high-skilled workers tend to be located in places that are easily accessible. Predictors related to the location of industrial activity (iron_steel, power, mining and cement) showed up as important for several member models but, perhaps in contradiction to expectations, were not uniquely related to the location of craft workers and operators and technicians and associate professionals, of which a large share is expected to live close to these facilities.

2.3 Relationship with spatial wealth

To evaluate the accuracy of our model estimates at grid level, we compared the predicted occupation shares with a high-resolution (2.4 km) global wealth index map (Chi et al., 2022). The wealth index is based on a machine learning model that used satellite imagery, mobile phone networks and connectivity data from Facebook as predictors and asset information from the Demographic and Health Surveys (DHS) as target variables, covering the period 2010-2018. The information in the DHS was collected by the U.S. Agency for International Development and is therefore independent of the population census that we used to train the super learners. Various studies showed that wealth is strongly correlated with income (e.g. Sahn et al., 2003), which, in turn, is determined for a large part by labor income, and, hence, type of occupation.

This relationship was confirmed by Figure 4, which shows that the proportion of low-skilled occupations, composed of agricultural workers ($R^2 = 0.43$), craft and operators ($R^2 = 0.46$), clerks and service workers ($R^2 = 0.27$) and elementary occupations ($R^2 = 0.29$), which make up 91% of the labor force (Figure S1), explained a considerable part of relative wealth at grid-cell level. Wealth tends to be higher in (semi-)urban areas and therefore is associated with lower (higher) shares of (non-)agricultural workers. There was a limited correlation between the wealth index and the share of managers and professionals, and technicians and associate professionals. A possible reason for this might be the limited coverage of (wealthy) high-skilled workers in the DHS. Another explanation could be the uneven distribution of wealth between high- and low-skilled workers within (urban) grid cells, which was not picked up by an average measure such as the wealth index.

3 Discussion

This paper presented an approach to downscale subnational labor statistics into high-resolution gridded maps using a super learner machine learning algorithm. The results showed that the predictions of the super learners were more accurate or at least comparable with those of the best-performing member models. We also found that none of six conventional machine learning algorithms showed consistent best-performance, implying that predictions using an ensemble approach are more robust than relying on a single algorithm for the analysis. Potential disadvantages of using ensemble approaches are the higher computational costs and complexity vis-à-vis fitting only a single type of machine learning model. Nonetheless, with the growing availability of high computing power, we expect this approach to be increasingly adopted by researchers with interest in the spatial prediction of social and biophysical indicators (e.g. Hengl et al., 2021).

This research can be extended, refined and improved in several directions. First, the predictive power of the super learners might be increased by collecting and incorporating additional spatial predictors. For example, the model for managers and professionals, which had relative poor validation metrics, might benefit from spatial layers that show the distance or travel time to schools, hospitals, city halls and other government buildings, which are typical locations of work for this type of occupation (see Stevens et al., 2015 for an example involving the distance to health facilities). Similarly, the accuracy of the craft workers and operators model might be improved by adding data on the location of industrial zones and factories. Unfortunately, digital maps with this type of information are not available for Vietnam. Overall, the predictive power of spatial indicators that measure the distance to work will probably be lower for high-skilled workers as they tend to commute over longer distances in comparison to low-skilled workers (O’Kelly et al., 2005; Sang et al., 2011).

Second, we aim to apply our approach to larger regions, such as South-East Asia and the world, resulting in a product that complements existing global maps with socio-economic indicators, such as population

Figure 4: Scatterplot of the relative wealth index from Chi et al. (2022) against the predicted occupation shares for (a) Managers and professionals, (b) Technicians and associate professionals, (c) Clerks and service workers, (d) Agricultural workers (e) Craft workers and operators and (f) Elementary occupations.

(www.worldpop.org), education (Graetz et al., 2019) and wealth (Chi et al., 2022). For a substantial number of countries, the required subnational labor statistics can be extracted from the population census, which is available by means of the IPUMS database (Minnesota Population Center, 2019). Another source of information is summary reports of national labor force surveys that are regularly published by national statistical agencies. The level of detail, however, can differ considerably between data sources and countries. In the case of Vietnam, we were able to obtain information for second-level administrative units. For many other countries, only a few data points, representing first-level administrative units or broad regions, might be available. Combining the data of a large number of countries will make it possible to train machine learning models that achieve higher accuracy and that can be used to generate plausible out-of-sample predictions for countries for which no data is available (Chi et al., 2022; Nicolas et al., 2016).

Finally, an interesting avenue of future research would be to investigate the possibility of implementing our approach to derive high-resolution maps for other (labor) indicators, such as gender-specific occupation maps. As mentioned above, differences in labor force participation between men and women have a strong spatial dimension, which can be disentangled by our analytical approach. Such an indicator would also inform spatial predictions of the proportion of women in managerial positions, which is one of the SDG5: Gender equality indicators. Another relevant indicator is the unemployment rate, listed under SDG8: Decent work and economic growth. Downscaling unemployment rates, however, is complex because it is the sum of frictional, cyclical, and structural components (Rissman, 1986). In advance, it is not clear which of these components will be captured by the various spatial predictors. For example, night lights are a proxy for economic activity (Gibson et al., 2021) and therefore might pick up the sum of all three components, while the distance to cropland, airports and industrial facilities might be correlated with structural unemployment rates. Although our analysis focused on downscaling labor data, it can equally be applied to disaggregate other statistics only available at the subnational level, including sectoral economic output, health and education information (Kalkuhl et al., 2021; Smits et al., 2019).

4 Materials and Methods

4.1 Target variables

Our primary target variable is the total number of workers within an occupation category. We distinguished between five broad occupation categories based on the most recent version of the International Standard Classification of Occupations (ISCO) (Table S1). The ISCO occupation classes are designed to capture two interrelated concepts: (a) the type of job since occupation is defined as a “set of jobs whose main task and duties are characterized by a high degree of similarity” and (b) skill level, which is defined as “function of complexity and range of tasks and duties to be performed in an occupation” (ILO, 2012). To estimate the number of workers O_i per occupation category $i = 1, \dots, 6$, we broke it down into three components:

$$O_i = WP \times LFPR \times OS_i \quad (1)$$

where WP is the working-age population, which, we defined as all persons aged 15 and older, in line with the standard ILO definition. $LFPR$ is the labor force participation rate, defined as the number of persons in the labor force as a percentage of the working-age population. The labor force is the sum of the number of persons employed and unemployed, aged 15 and older, which, by definition, is equal to the sum of all workers with an occupation. OS is the share of the labor force with occupation i .

Our main source of information was the 2009 Population and Housing Census organized by the Vietnam General Statistics Office that is available from IPUMS (Minnesota Population Center, 2019). Representative labor statistics were available for 674 districts, which are the second-level administrative units in Vietnam. The census included questions on the type of occupation based on the detailed ISCO 08 classification. We used this information to calculate the share of each major occupation group at the district level. Similarly, we also used the census to estimate the district labor force participation rate by dividing all persons that are categorized as being in the labor force by the total population aged 15 and older.

To determine the working age for each grid cell, we used population maps with 5-year age group compositions from Pezzulo et al. (2017). These maps were created by combining a gridded population map for Vietnam (Stevens et al., 2015) with subnational data on the shares of age groups to estimate grid cell values. The main data source for these products was the 2009 Vietnam population census, which we also used as the main input.

As a final step, we used a logit transformation to change the scale of our dependent variables, which are both proportions that are bounded between zero and one. As the model predictions are not guaranteed to be within this range, such a transformation is recommended to change the scale to between negative and positive infinity (Kuhn & Johnson, 2020).

4.2 Geospatial predictors

We selected 32 predictors, which we expected to explain the observed spatial distribution of the labor force participation rate and occupation categories (Table S2). The majority of predictors were taken from the WorldPop (worldpop.org) and WorldClim (worldclim.org) open access archives, which contain a diverse range of remote sensing and other geospatial data at a resolution of 3-30 arc-seconds that cover multiple periods, including land cover, night lights, transport networks, topography and climate indicators (Figure S3-S8). In addition, we added geospatial layers with information on distance to large industrial facilities (energy, iron, steel and cement plants, and mines) and major transport hubs (airports and ports), which are often part of industrial zones, where a large number of people are employed (Figure S7 and S9). As workers tend to live in the vicinity of their work, we assumed that these layers have predictive power and support the training of the machine learning models. Information on the location of power, iron and steel and cement plants was taken from the Global Infrastructure Emissions (GID) database (www.gidmodel.org.cn). The GID contains information on emissions from energy-intensive industries, which can be regarded as a proxy for their location. In addition, we collected information on the location of mining areas, airports and ports. These datasets were further processed to create geospatial layers with distance information.

Where possible, we selected layers with information for the year 2009, consistent with the population census data. However, in a few cases, in particular for the predictors related to the location of industrial facilities, only recent data was available. To harmonize the data, all raster layers were resampled to a resolution of 30 arc seconds in WGS 84 projection. As the occupation data is available at the second administrative unit level, all predictors were aggregated to the same level using the median values for each administrative area.

We applied three feature engineering steps that result in better performance of machine learning models (Kuhn & Johnson, 2020): (a) We applied a Yeo-Johnson power transformation (Yeo et al., 2000) to make the variable distributions more symmetric, (b) we normalized all data to have a standard deviation of one and a mean of zero, where mean values and standard deviations were estimated from the training dataset and (c) we removed all variables that had a correlation of 0.9 or larger with other variables.

4.3 Super learner

The process to train the super learner followed the steps as described in Laan et al. (2007). In total, we trained seven separate super learners; one model to spatially predict the labor force participation rate (*LFPR*) and six models to predict the shares for each occupation class (*OS_i*). Each super learner combined the results of six different machine learning algorithms that have been used extensively in (spatial) prediction exercises: random forest, (`random_forest`) extreme gradient boosting (`xgboost`), neural network (`neural_network`), polynomial (`svm_poly`) and radial (`svm_radial`) basis support vector machines, and generalized linear model via penalized maximum likelihood (`glmnet`). All super learner member models were trained using data at the second-level administrative unit (674 observations).

All models (i.e. super learner and member models) were implemented using the `tidymodels` framework (Kuhn & Wickham, 2020) in the R software environment (R Core Team, 2021). We started by applying the `tidymodels` `tune` and `dials` packages to optimize the tuning of hyperparameters for all super learner member models. For each model, we conducted a grid search with 30 model-specific parameter combinations that were selected using a Latin hypercube design. Each parameter combination was evaluated and fitted by means of 5 repeats of 10 fold cross-validation, using 80% of the dataset for training. Inspection revealed that a few models had very poor performance, predicting (near) constant values. Therefore, we removed all models for which the predictions had a standard deviation lower than 0.001. We preferred to exclude these poor models as it results in parsimonious super learner models but with the same performance. In the next step, we tuned a regularized generalized linear model to optimally combine the predictions of the member models and determine the relative weights for the super learner. We used the default settings that constrain the coefficients of the super learner to be non-negative. We analyzed the accuracy of our modeling framework by fitting all selected member models (i.e. those with a weight larger than zero) and the super learner on hold-out data that contains 20% of the main dataset. For each model, we derived the RMSE and the R^2 . The super learners were used to predict the labor force participation rates and occupation shares for all grid cells in Vietnam. We applied an inverse logit transformation to obtain values between zero and one. For the predictions of the six occupation shares we used a standardization function to ensure that the total sum of shares equals one in each grid cell:

$$occupation_{ij}^c = \frac{occupation_{ij}}{\sum_{i=1}^6 occupation_{ij}} \quad (2)$$

where $occupation_{ij}^c$ is the corrected share for occupation i in grid cell j . Finally, the occupation maps were produced by multiplying grid-level information on labor force participation rate, occupation share and working-age population. We followed the approach used by Hengl (2021) to calculate grid-cell prediction errors for super learner models. This involved two steps (a) fitting a quantile regression random forest model using the training dataset predictions of the super learner member as input and (b) estimating error statistics per grid cell using the `forestError` package that implements the method proposed by Lu et al. (2021).

For each member model we created variable importance measures that demonstrate how much a models' performance (measured by the RMSE) changes if an explanatory variable is removed, as well as accumulated local effects plots, which show how features relate to the machine learning predictions on average. The methods

to create both types of plots, implemented with the R DALEX package (Biecek, 2018), are model-agnostic and do not assume anything about the structure of the super learner member models (Biecek et al., 2021).

5 Data Availability.

The main results and input datasets of this study are presented in the Supplementary Information. All input and output maps produced for this study can also be accessed by means of an interactive web application: <https://shiny.wur.nl/occupation-map-vnm>.

6 Acknowledgments

This research was funded by a grant from Wageningen University (KB35-005-001) and a contribution from the Wageningen University and Research investment fund. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We would like to thank Tom Hengl for advice on how to calculate the prediction errors.

References

- Autor, D. H., Dorn, D., & Hanson, G. H. (2013). The China Syndrome: Local Labor Market Effects of Import Competition in the United States. *American Economic Review*, 103(6), 2121–2168. <https://doi.org/10.1257/aer.103.6.2121>
- Biecek, P. (2018). DALEX: Explainers for Complex Predictive Models in R. *Journal of Machine Learning Research*, 19(84), 1–5. Retrieved from <http://jmlr.org/papers/v19/18-416.html>
- Biecek, P., & Burzykowski, T. (2021). *Explanatory Model Analysis*. New York: Chapman; Hall/CRC.
- Chi, G., Fang, H., Chatterjee, S., & Blumenstock, J. E. (2022). Microestimates of wealth for all low- and middle-income countries. *Proceedings of the National Academy of Sciences*, 119(3), e2113658119. <https://doi.org/10.1073/pnas.2113658119>
- Duernecker, G., & Herrendorf, B. (2021). *Structural Transformation of Occupation Employment*. Retrieved from <https://ssrn.com/abstract=3932029>
- Ellison, G., & Glaeser, E. L. (1999). The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration? *American Economic Review*, 89(2), 311–316. <https://doi.org/10.1257/AER.89.2.311>
- Giattino, C., Ortiz-Ospina, E., & Roser, M. (2013). Working Hours. *Our World in Data*. Retrieved from <https://ourworldindata.org/working-hours>
- Gibbons, R., Katz, L. F., Lemieux, T., & Parent, D. (2005). Comparative Advantage, Learning, and Sectoral Wage Determination. *Journal of Labor Economics*, 23(4), 681–724. <https://doi.org/10.1086/491606>
- Gibson, J., Olivia, S., Boe-Gibson, G., & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, 149, 102602. <https://doi.org/10.1016/J.JDEVECO.2020.102602>
- Graetz, N., Woyczynski, L., Wilson, K. F., Hall, J. B., Abate, K. H., Abd-Allah, F., ... Hay, S. I. (2019). Mapping disparities in education across low- and middle-income countries. *Nature*. <https://doi.org/10.1038/s41586-019-1872-1>
- Graff Zivin, J., & Neidell, M. (2014). Temperature and the Allocation of Time: Implications for Climate Change. *Journal of Labor Economics*, 32(1), 1–26. <https://doi.org/10.1086/671766>
- Hengl, T. (2021). *Extrapolation is tough for trees (tree-based learners), combining learners of different type makes it less tough*. Retrieved from <https://medium.com/nerd-for-tech/extrapolation-is-tough-for-trees-tree-based-learners-combining-learners-of-different-type-makes-659187a6f58d>

- Hengl, T., Miller, M. A. E., Križan, J., Shepherd, K. D., Sila, A., Kilibarda, M., ... Crouch, J. (2021). African soil properties and nutrients mapped at 30 m spatial resolution using two-scale ensemble machine learning. *Scientific Reports*, 11(1), 6130. <https://doi.org/10.1038/s41598-021-85639-y>
- ILO. (2012). *International Standard Classification of Occupations. Volume 1: Structure, group definitions and correspondence tables*. Geneva, Switzerland: International Labour Office.
- ILO. (2021). *World Employment and Social Outlook – Trends 2021*. Geneva, Switzerland: International Labour Office.
- Kalkuhl, M., Kotz, M., & Wenz, L. (2021). *DOSE - The MCC-PIK Database Of Subnational Economic output*. <https://doi.org/10.5281/ZENODO.4681306>
- Karabarbounis, L., & Neiman, B. (2014). The Global Decline of the Labor Share. *The Quarterly Journal of Economics*, 129(1), 61–103. <https://doi.org/10.1093/QJE/QJT032>
- Krugman, P. (1991). *Geography and Trade*. Cambridge, MA: MIT Press.
- Kuhn, M., & Johnson, K. (2020). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. CRC Press.
- Kuhn, M., & Wickham, H. (2020). *Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles*. Retrieved from <https://www.tidymodels.org>
- Laan, M. J. van der, Polley, E. C., & Hubbard, A. E. (2007). Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1). <https://doi.org/10.2202/1544-6115.1309>
- Lee, B., Binns, T., & Dixon, A. (2010). The Dynamics of Urban Agriculture in Hanoi, Vietnam. *Field Actions Science Reports. The Journal of Field Actions*, (Special Issue 1), 0–8.
- Leyk, S., Gaughan, A. E., Adamo, S. B., De Sherbinin, A., Balk, D., Freire, S., ... Pesaresi, M. (2019). The spatial allocation of population: a review of large-scale gridded population data products and their fitness for use. *Earth System Science Data*, 11(3), 1385–1409. <https://doi.org/10.5194/ESSD-11-1385-2019>
- Lima, C. Z. de, Buzan, J. R., Moore, F. C., Baldos, U. L. C., Huber, M., & Hertel, T. W. (2021). Heat stress on agricultural workers exacerbates crop impacts of climate change. *Environmental Research Letters*, 16(4), 44020. <https://doi.org/10.1088/1748-9326/abeb9f>
- Lu, B., & Hardin, J. (2021). A Unified Framework for Random Forest Prediction Error Estimation. *Journal of Machine Learning Research*, 22(8), 1–41. Retrieved from <http://jmlr.org/papers/v22/18-558.html>
- Mammen, K., & Paxson, C. (2000). Women’s Work and Economic Development. *Journal of Economic Perspectives*, 14(4), 141–164. <https://doi.org/10.1257/jep.14.4.141>
- Minnesota Population Center. (2019). *Integrated Public Use Microdata Series, International: Version 7.2 [dataset]*. <https://doi.org/https://doi.org/10.18128/D020.V7.2>
- Nicolas, G., Robinson, T. P., Wint, G. R. W., Conchedda, G., Cinardi, G., & Gilbert, M. (2016). Using Random Forest to Improve the Downscaling of Global Livestock Census Data. *PLOS ONE*, 11(3), e0150424. <https://doi.org/10.1371/journal.pone.0150424>
- O’Kelly, M. E., & Lee, W. (2005). Disaggregate Journey-to-Work Data: Implications for Excess Commuting and Jobs–Housing Balance. *Environment and Planning A: Economy and Space*, 37(12), 2233–2252. <https://doi.org/10.1068/a37312>
- Orlov, A., Sillmann, J., Aunan, K., Kjellstrom, T., & Aaheim, A. (2020). Economic costs of heat-induced reductions in worker productivity due to global warming. *Global Environmental Change*, 63, 102087. <https://doi.org/10.1016/j.gloenvcha.2020.102087>
- Osgood-Zimmerman, A., Millea, A. I., Stubbs, R. W., Shields, C., Pickering, B. V., Earl, L., ... Hay, S. I. (2018). Mapping child growth failure in Africa between 2000 and 2015. *Nature*, 555(7694), 41–47. <https://doi.org/10.1038/nature25760>

- Pezzulo, C., Hornby, G. M., Sorichetta, A., Gaughan, A. E., Linard, C., Bird, T. J., ... Tatem, A. J. (2017). Sub-national mapping of population pyramids and dependency ratios in Africa and Asia. *Scientific Data*, 4(1), 170089. <https://doi.org/10.1038/sdata.2017.89>
- Pokhriyal, N., & Jacques, D. C. (2017). Combining disparate data sources for improved poverty prediction and mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 114(46), E9783–E9792. <https://doi.org/10.1073/pnas.1700319114>
- Polley, E. C., & Laan, M. J. van der. (2010). *Super Learner In Prediction*.
- R Core Team. (2021). *R: A Language and Environment for Statistical Computing*. Retrieved from <https://www.r-project.org/>
- Rickard, S. J. (2020). Economic Geography, Politics, and Policy. *Annual Review of Political Science*, 23(1), 187–202. <https://doi.org/10.1146/annurev-polisci-050718-033649>
- Rissman, E. R. (1986). What is the natural rate of unemployment? *Federal Reserve Bank of Chicago Economic Perspectives*, 10, 3–17.
- Sahn, D. E., & Stifel, D. (2003). Exploring Alternative Measures of Welfare in the Absence of Expenditure Data. *Review of Income and Wealth*, 49(4), 463–489. <https://doi.org/10.1111/J.0034-6586.2003.00100.X>
- Sang, S., O’Kelly, M., & Kwan, M.-P. (2011). Examining Commuting Patterns. *Urban Studies*, 48(5), 891–909. <https://doi.org/10.1177/0042098010368576>
- Smits, J., & Permanyer, I. (2019). The Subnational Human Development Database. *Scientific Data* 2019 6:1, 6(1), 1–15. <https://doi.org/10.1038/sdata.2019.38>
- Stevens, F. R., Gaughan, A. E., Linard, C., Tatem, A. J., Jarvis, A., & Hashimoto, H. (2015). Disaggregating Census Data for Population Mapping Using Random Forests with Remotely-Sensed and Ancillary Data. *PLOS ONE*, 10(2), e0107042. <https://doi.org/10.1371/journal.pone.0107042>
- Tarp, F. (Ed.). (2017). *Growth, Structural Transformation, and Rural Change in Viet Nam*. <https://doi.org/10.1093/acprof:oso/9780198796961.001.0001>
- Yeo, I.-K., & Johnson, R. A. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika*, 87(4), 954–959.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [si.pdf](#)