2023 Joint Urban Remote Sensing Event (JURSE) | 978-1-6654-9373-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/JURSE57346.2023.10144145

# LSTM models for spatiotemporal extrapolation of population data

Christian Geiß, Jana Maier, Emily So, and Yue Zhu

German Remote Sensing Data Center (DFD), German Aerospace Center (DLR) 82234 Wessling-Oberpfaffenhofen, Germany; christian.geiss@dlr.de, jana.maier@dlr.de,
Centre for Risk in the Built Environment, University of Cambridge, Cambridge, UK;
ekms2@cam.ac.uk, yz591@cam.ac.uk

*Abstract*— **The anticipation of future geospatial population distributions is crucial for numerous application domains. Here, we capitalize upon existing gridded population time series data sets, which are provided on an open source basis globally, and implement a machine learning model tailored for time series analysis, i.e., Long Short Term Memory (LSTM) network. In detail, we harvest WorldPop population data and learn an LSTM model for anticipating population along a three-year interval. Experimental results are obtained from Peru's capital Lima, which features a high population dynamic. To gain insights regarding the competitive performance of LSTM models in this application context, we also implement multilinear regression and Random Forest models for comparison. The results underline the usefulness of temporal models, i.e., LSTM, for forecasting gridded population data.**

***Index Terms*— spatiotemporal population modeling; time series data; LSTM models; Lima, Peru**

## I. Introduction

Geospatial modeling of the population is crucial for numerous application domains such as natural hazard risk assessment [1], accessibility assessment of medical support [2], and general monitoring of the progress towards development goals [3], among others. The dynamic change of population distributions due to population growth and urbanization processes [4] induces the need to constantly update and eventually anticipate future geospatial population distributions.

To anticipate future geospatial population distributions, various techniques can be considered generally: Rule-based methods establish a set of explicitly defined rules for transition trajectories over time. This family of methods contains i) Cellular Automata techniques [5] which represent discrete spatiotemporal dynamic systems based on local rules; ii) Agent-based Modelling which simulates dynamic interactions among agents in a virtual environment [6]; iii) Markov Chain Models which represent a stochastic process that produces sequential states in which each prediction is dependent on the previous state [7].

Techniques of empirical inference were also utilized for predicting transition trajectories in the context of population modeling. The underlying idea is to infer a decision rule (e.g., a function) from limited but properly encoded prior knowledge (i.e., labeled training samples). For instance, Chen *et al.* [8] integrate high-resolution historical population maps and multiple machine learning models, i.e., XGBoost, Random Forest (RF), and Neural Network, to predict future built-up land and population distributions. Kubota *et al.* [9] implemented a Graph Convolutional Network for short-term population prediction based on population count data collected through mobile phones. Zheng and Zhang [10] implement a Convolutional LSTM (ConvLSTM) network for weekly population distribution prediction based on geolocated social media data, i.e., Tencent positioning data.

In contrast to previous works, to alleviate the frequently costly compilation of training data, here we capitalize upon existing gridded population time series data sets, which are provided on an open source basis globally, and implement a machine learning model tailored for time series analysis, i.e., Long Short-Term Memory (LSTM) network [11]. Different initiatives offer continuous gridded geospatial population data over a long time frame: WorldPop [3],[12], and LandScan [13] provide yearly geospatial population estimates starting in the year 2000. The data sets are created with a top-down approach by disaggregating census information based on satellite imagery and ancillary spatial covariates.

For this study, we uniquely harvest WorldPop population data and learn an LSTM model for anticipating population along a three-year interval. Experimental results are obtained from Peru's capital Lima, which features a high population dynamic. To gain insights regarding the competitive performance of LSTM models in this application context, we also implement multilinear regression (MLR) and RF models for comparison.

The remainder of the paper is organized as follows. In **Section 2** we detail the proposed methodology. We describe the study area and experimental setup in **Section 3**. Experimental results are revealed in **Section 4** and concluding remarks are given in **Section 5**.

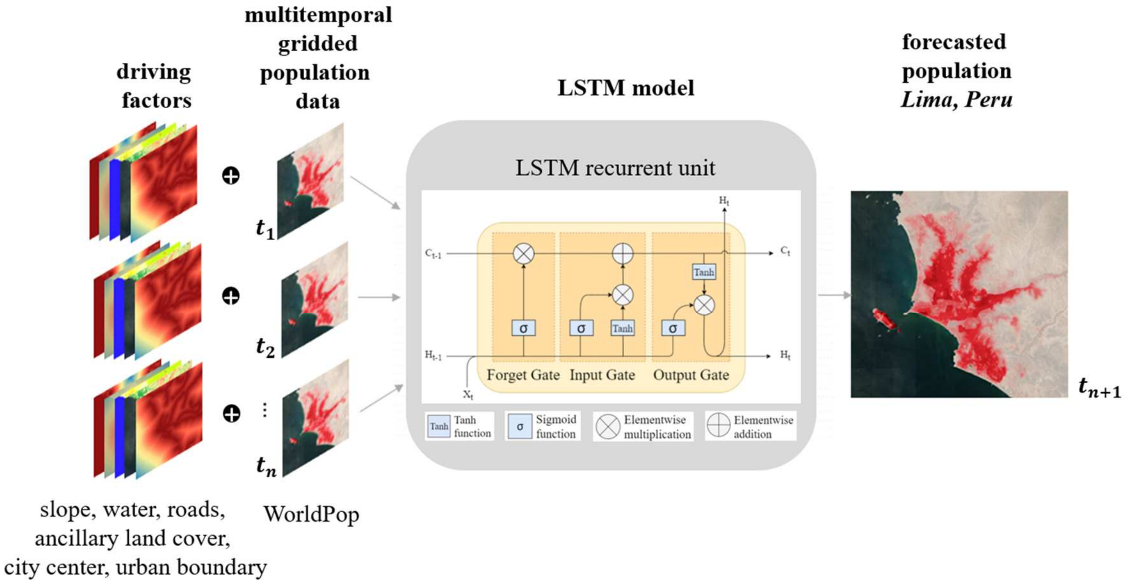


**Fig. 1. Overview of the workflow for spatiotemporal forecasting of population data**

## II. Methodology

Fig. 1 provides an overview of the proposed workflow for spatiotemporal forecasting of population data. First, we compute a set of geospatial covariates, i.e., driving factors. Subsequently, multitemporal gridded population data are compiled. The population data of time steps $t_1, t_2, \ldots, t_n$ and the corresponding driving factors are concatenated as the input for the LSTM, which maps the input to a prediction of the population at time step $t_{n+1}$. The main model architecture comprises an LSTM recurrent unit with:

$$i_t = \sigma(W_{xi}X_t + W_{hi}H_{t-1} + W_{ci} \circ C_{t-1} + b_i) \qquad (1)$$

$$f_t = \sigma(W_{xf}X_t + W_{hf}H_{t-1} + W_{cf} \circ C_{t-1} + b_f)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_{xc}X_t + W_{hc}X_{t-1} + b_c)$$

$$o_t = \sigma(W_{xo}X_t + W_{ho}H_{t-1} + W_{co} \circ C_t + b_o)$$

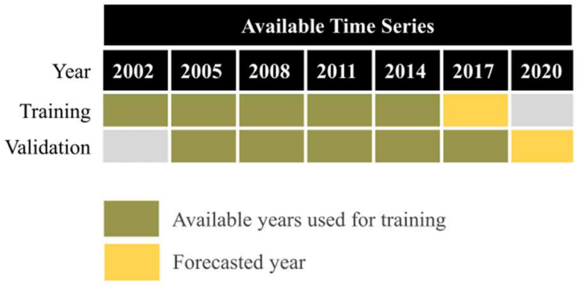
$$H_t = o_t \circ \tanh(C_t)$$

where $X_t$ represents the input to the cell, $C_t$ the memory state, and $H_t$ the hidden state. The notation '∘' denotes the Hadamard product or element-wise product. In the equations, $i_t$, $f_t$, and $o_t$ refer to the input, forget, and output gates, respectively, $t$ is the time-step, $\sigma$ the sigmoid activation function, tanh the hyperbolic tangent function, and $W$ are the weight matrices and $b$ the biases, respectively.

## III. Data Sets and Experimental Setup

The study area comprises the settlement area of Peru's capital Lima with a spatial coverage of approximately 6500 square kilometers. The data set consists of yearly multi-temporal gridded population data with a spatial resolution of 100 meters from WorldPop [3],[12] for the period 2000-2020 and land change driving factors. The latter includes (1) slope, (2) distance to water, (3) distance to roads, (4) ancillary land cover, (5) distance to the city center, and (6) distance to the urban boundary. The slope was calculated from the Copernicus Digital Elevation Model. The data source for computing road distances was extracted from OpenStreetMap. To compute the distance to water, water bodies from the Copernicus layer were combined with waterways from OpenStreetMap.

The data set is split into training data set and validation data set along the temporal dimension. The training data set contains earlier six time steps (2002, 2005, …, 2017), whereas the validation dataset contains later six time steps (2005, 2008, …, 2020). In both training and validation data sets, the variables of the first five time steps were adopted as input and the last time step was used as the ground truth labels. As such, the target of the training data set is to predict the population of the year 2017, and the goal of the validation dataset is to forecast the population map for the year 2020 (Fig 2).



| | Available Time Series | | | | | | |
|---|---|---|---|---|---|---|---|
| Year | 2002 | 2005 | 2008 | 2011 | 2014 | 2017 | 2020 |
| Training | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟨 | ⬜ |
| Validation | ⬜ | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟨 |

🟩 Available years used for training

🟨 Forecasted year

**Fig. 2. training/validation concept**

All the tested models were trained for 50 epochs, the optimizer was Adam, the loss function was mean squared error loss, and the initial learning rate was set to 0.0012 and was reduced by the factor 0.1 through a learning rate scheduler, when the error reached a minimum plateau. To evaluate the proposed framework, two baseline methods were adopted, i.e., MLR and RF. Thereby, the hyperparameters of RF were tuned as follows: ntree = 500 and mtry = 1,2,⋯,51.

## IV. Experimental Results

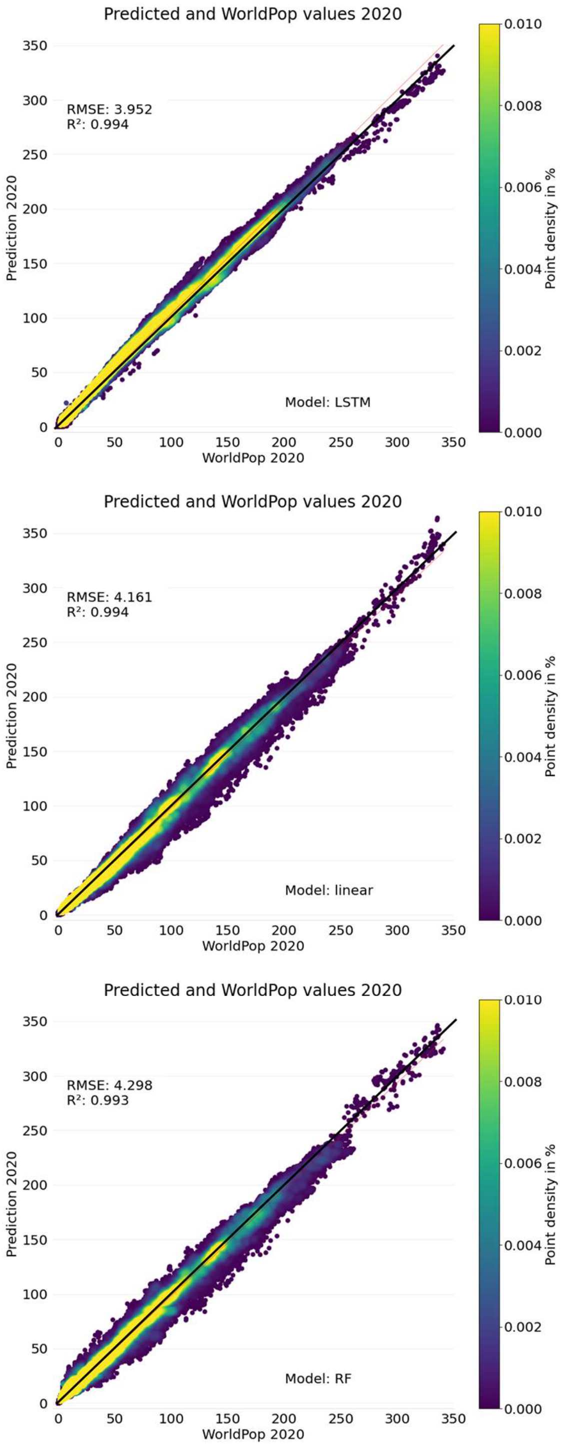


**Fig. 3. Scatter plots and error measures for the considered methods**

To provide a first comparative overview, Fig. 3 contains scatter plots of the methods for the predicted year 2020. It reveals that all models feature a strong concentration of the density along the one-to-one line. However, the uncertainty in terms of RMSE could be reduced from 4.298 (RF) and 4.161 (MLR), respectively, to 3.952 (LSTM) while maintaining an excellent model fit (R = 0.994).

Fig 4a shows the ground truth and corresponding model estimates which reflect the spatially strongly varying population distribution in Lima. Thereby, abrupt changes in population numbers along inner-city administrative boundaries can be observed, while a continuous decrease of the population along the border of the settlement body is traceable. Fig. 4b captures the momentum, i.e., population change between 2017 and 2020. Thereby, the LSTM can provide pronounced change patterns which, however, exceed the reference with respect to magnitude. In contrast, both MLR and RF are hardly able to capture the change of dynamic areas properly. Nevertheless, all models reflect areas of dominantly decreasing (blue) and increasing (red) population numbers.

Finally, to visualize actual differences in the predictions regarding the reference population distribution, Fig 4c provides prediction differences to the actual numbers of 2020. Thereby, it can be traced that the LSTM-based predictions overestimate population numbers, while both the MLR-based and RF-based predictions underestimate population numbers for a majority of areas (also revealed by the regression line in Fig. 3).

## V. Conclusions and Outlook

This study underlines the usefulness of temporal models, i.e., LSTM, for forecasting of gridded population data. In the future, we aim to equip LSTM with a bidirectional learning mechanism, i.e., running the model inputs in two ways, one from past to future and one from future to past, and also implement and evaluate ConvLSTM models [14] that are dedicated to process spectral-spatial sequential data.

## Acknowledgements

This study has been conducted as part of the projects RIESGOS 2.0 (03G0905A-B), funded by the German Federal Ministry of Education and Research (BMBF).

## References

[1] Geiß, C., Priesmeier, P., Aravena Pelizari, P., Soto, A., Schöpfer, E., Riedlinger, T., Villar Vega, M., Santa Maria, H., Gomez Zapata, C., Pittore, M., So, E., Fekete, A., and Taubenböck, H. (2022): Benefits of Global Earth Observation Missions for Disaggregation of Exposure Data and Earthquake Loss Modelling – Evidence from Santiago de Chile. *Natural Hazards*. https://doi.org/10.1007/s11069-022-05672-6

[2] Rauch, S., Taubenböck, H., Knopp, and C., Rauh, J. (2021): Risk and space: modelling the accessibility of stroke centers using day- & nighttime population distribution and different transportation scenarios. *Int J Health Geogr.*, 20(31).

[3] Lloyd, C. T., Sorichetta, A., and Tatem, A. J. (2017): High resolution global gridded data for use in population studies. *Scientific Data*, 4(1).

[4] UN Habitat. (2016). Urban Impact, (06). UN Habitat, United Nations Human Settlements Programme.

[5] Clarke, K. (2014): Cellular Automata and Agent-Based Models. Fischer, M.M., Nijkamp P. (eds.): Handbook of Regional Science, pp. 1217–1233, Springer – Berlin, Heidelberg.

[6] Abar, S., Theodoropoulos, G.K., Lemarinier, P., and O'Hare, G. M. P. (2017): Agent Based Modelling and Simulation Tools: A Review of the State-of-Art Software. *Computer Science Review*, 24, 13–33.

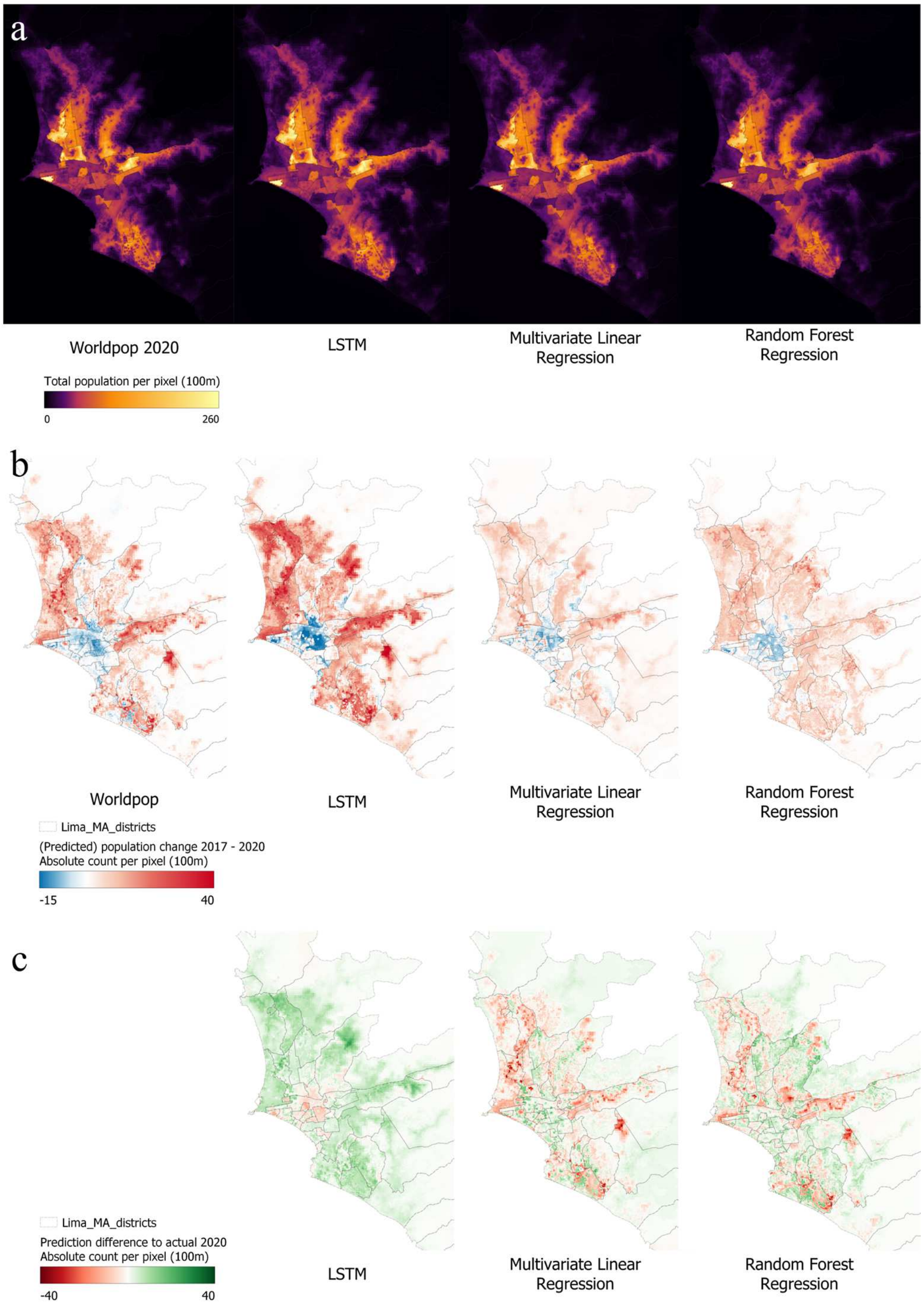

a
Worldpop 2020
LSTM
Multivariate Linear Regression
Random Forest Regression
Total population per pixel (100m)
0
260
b
Worldpop
LSTM
Multivariate Linear Regression
Random Forest Regression
Lima_MA_districts
(Predicted) population change 2017 - 2020
Absolute count per pixel (100m)
-15
40
c
Lima_MA_districts
Prediction difference to actual 2020
Absolute count per pixel (100m)
-40
40
LSTM
Multivariate Linear Regression
Random Forest Regression

**Fig. 4. Prediction maps and corresponding error maps for the considered methods**

[7] Gagniuc, P.A. (2017): Markov chains: from theory to implementation and experimentation. John Wiley & Sons.

[8] Chen, Y., Li, X., Huang, K., Luo, M., and Gao, M. (2020): High-Resolution Gridded Population Projections for China Under the Shared Socioeconomic Pathways. *Earth's Future*, 8(6).

[9] Kubota, Y., Ohira, Y., and Shimizu, T. (2022): Attention-based Contextual Multi-View Graph Convolutional Networks for Short-term Population Prediction. Proceedings of UrbComp Workshop '21: ACM SIGSPATIAL (UrbComp Workshop '21).

[10] Zheng, Z., and Zhang, G. (2020): The prediction of finely-grained spatiotemporal relative human population density distributions in China. *IEEE Access*, 8, 181537.

[11] Hochreiter, S., and Schmidhuber, J. (1997): Long Short-Term Memory. *Neural Computation* 9(8): 1735-1780.

[12] Stevens, F. R., Gaughan, A. E., Linard, C., and Tatem, A. J. (2015): Disaggregating Census Data for Population Mapping Using Random Forests with Remotely Sensed and Ancillary Data. *PLOS ONE*, 10(2), e0107042.

[13] Dobson, J.E.; Bright, E.A.; Coleman, P.R., Durfee, R.C., and Worley, B.A. (2000): LandScan: A Global Population Database for Estimating Populations at Risk. *Photogrammetric Engineering & Remote Sensing*, 66(7), 849–857.

[14] Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-k., and Woo, W.-c. (2015). Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in Neural Information Processing Systems*, volume 28.