UC Santa Barbara
GIScience 2021 Short Paper Proceedings

Title
Spatially-explicit forecasting of racial change

Permalink
https://escholarship.org/uc/item/4n31h85w

Authors
Stepinski, Tomasz
Dmowska, Anna

Publication Date
2021-09-01

DOI
10.25436/E29G6Q

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Spatially-explicit forecasting of racial change

Tomasz f. Stepinski[1]
Space Informatics Lab, Department of Geography and GIS, University of Cincinnati, Cincinnati, OH, USA

Anna Dmowska
Institute of Geoecology and Geoinformation, Adam Mickiewicz University, Poznan, Poland

Abstract

Spatio-racial distributions in major US cities change on the timescale of a single decade. Here we describe a methodology to forecast such changes a decade ahead. First, we transform the data from population counts to a grid of categorical population types. Then, we build an empirical model of past change using supervised machine learning and extrapolate it into the future to make a prediction. The model uses only statistics of population categories as features, there are no ancillary variables. To account for the non-stationarity of the change we use a synthetic training dataset based on past transitions and estimated future frequencies of these transitions. The methodology is described and validated by training a model on 1990-2000 data and using it to predict spatio-racial distributions in 2010. This prediction is then compared to the actual spatio-racial 2010 distribution. We have found that a highly accurate model of change can be constructed using this methodology. Extrapolating such models to the future results in some loss of accuracy, but the method still yields satisfactory predictions.

1 Introduction

Demographic predictions help to plan for the future and provide a priori information on a variety of problems to decision-makers and other stakeholders. Predictions at the country level, broken down by age and sex, are published by the UN every two years. For the US there are also county-level predictions by age, sex, and race [4]. Predictions on a small scale - called spatially-explicit predictions (SEP) have also been published [11, 7, 18]. SEPs for population counts and racial composition are important because they can inform and affect local socioeconomic planning [14]. In particular, spatially-specific predictions of racial change could be used for informed allocation of educational, housing, and transportation resources [16, 1, 17].

To address this issue we have developed an empirical SEP methodology for forecasting future racial spatial distribution using supervised machine learning [15]. Racial geography in a multiracial city is best conceptualized by the spatio-temporal racial probability distribution (STRD) $p(x, y, r; t)$, where x and y are spatial coordinates, r is a race variable, and t is a time. Because STRD is complex and difficult to work with, we transform STRD from the point pattern to the raster of small areal units. Furthermore, we classify these units into a finite number of population categories based on within-the-unit racial shares. Such approximation

[1] Corresponding author

preserves information about local diversity but loses information about population density; the STRD becomes a racial map - a raster of categorical values. As such, its form resembles the form of land cover maps, a domain where change forecasting has been done for decades. Although our method is inspired by land change methods [9, 12, 13], it uses a novel technical approach appropriate for our context.

The basic concept is to "learn" SEP model data from two consecutive census years in the past and apply this model to predict racial distribution for the upcoming census year. We use only statistics of the racial data as learning "features". No ancillary ("change drivers") variables are used. Observation of a trend of racial change in major US cities over several decades indicates that forecasting can be made without resorting to driver variables. Based on this observation we postulate that the category of a focus unit in the next census year depends on its present category and on statistics of other units' categories in its neighborhood. The specific function that assigns the future category of the focus unit is learned using a machine learning algorithm. This function is an empirical model of racial change in a given decade. Assuming that the next decade has racial dynamics similar to that in the decade used to learn the model, the model can be used for forecasting. Finally, we use a synthetic training dataset to account for the non-stationarity of change dynamics. Results are shown for Cook County, IL which includes the city of Chicago.

2 Data and Methods

The input data are 300 m resolution racial grids covering Cook County in 1990, 2000, and 2010 – the result of regridding finer 30 m resolution grids provided by SocScape (http://www.socscape.edu.pl); see [2] for the description. We use gridded data for computational convenience. We consider five subpopulations: White, Black, Asian, Hispanics, and others. Grid cells – our areal units – are not necessarily racially homogeneous. Therefore,we have classified them into 11 population categories [5]: (White, low diversity (WL), Black, low diversity (BL), Hispanics, low diversity (HL), Asian, low diversity (AL), others, low diversity OL, White, medium diversity (WM), Black, medium diversity (BM), Hispanics, medium diversity (HM), Asian, medium diversity (AM), others, medium diversity (OM), high diversity (Hdiv)).

An artificial neural network (ANN) algorithm is used to build a model of racial change. We use the architecture called the Self-Normalizing Net (SNN) [8]. Features are shares of cells' categories in the neighborhood of a focus areal unit in the proceeding year and a target is a category of this unit in the succeeding year. Training data are records for each unit, each record consists of values of features and a value of a target. Passing training data for all areal units through training constitutes one round. Training a model over multiple rounds increases its accuracy as measured on the basis of the performance on the training set. To prevent overfitting the algorithm monitors the performance of a model over the rounds on a held-out validation dataset. The algorithm stops rounds when the performance on the validation set stops improving. For our calculations, we used implementations of the SNN algorithm in the Wolfram Research Mathematica [10].

To account for changes in racial dynamics and improve the quality of forecasting we introduce a synthetic training set. Different racial dynamic does not necessarily mean different features $\rightarrow$ target records (transitions) but it means different frequencies of these transitions. Thus, the synthetic training dataset uses past transition records but with their frequencies estimated for transition ending in the year of prediction. First, we estimate the composition of categories in a year of prediction using the linear extrapolation from 1990 and

2000 compositions. Next, we estimate the number of transitions during the prediction decade using the Maximum Entropy Principle (MaxEnt) [6]. We used the function FindMaximum in the Wolfram Research Mathematica [10] to solve the constrained maximization problem required by the MaxEnt.

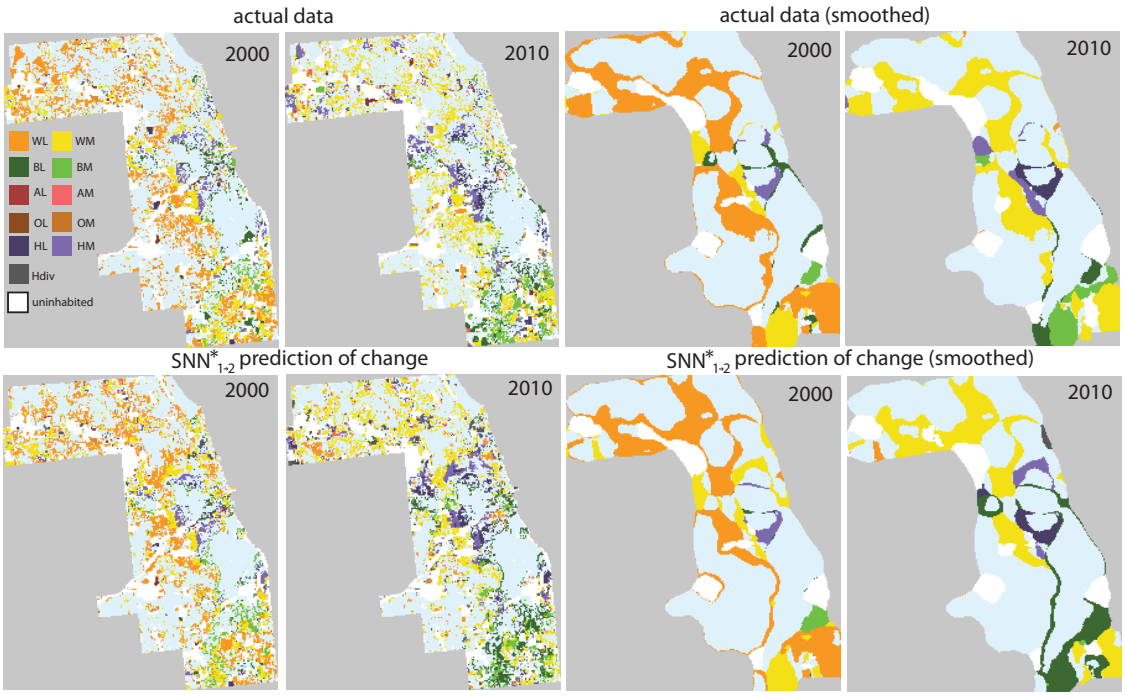

Figure 1 Changes in values of cells' categories between 2000 and 2010; actual changes (upper row) and change prediction by the $SNN^{\star}_{1\to 2}$ model (lower row). Light blue color indicates cells that did not change their category between 2000 and 2010.

3 Results

We calculated three different models, all for Cook County, IL. The $SNN_{2\to 3}$ is a model trained on 2000 and 2010 data and applied to 2000 data to predict racial distribution in 2010. The role of this model is to check whether our principle of racial change, as stated in the Introduction, can translate into an accurate model of racial change. Its classification accuracy is 0.82 and the visual inspection indicates that predicted and actual racial maps in 2010 are practically indistinguishable which validates the principle of change. The $SNN_{1\to 2}$ is a model trained on 1990 and 2000 data and applied to 2000 data to predict racial distribution in 2010. This model extrapolates the racial dynamics from the 1990-2000 decade into 2000-2010 decade. Its classification accuracy is 0.64; the visual inspection reveals noticeable differences between predicted and actual racial maps in 2010. This is because the racial dynamics has changed and extrapolation alone cannot account for all changes. The $SNN^{\star}_{1\to 2}$ model is trained on the synthetic dataset (that uses 1990-2000 transition records and 2000-2010 transition frequencies) to predict racial distribution in 2010. The classification accuracy is 0.63 (no improvement over the $SNN_{1\to 2}$ model) but the visual inspection shows significantly more similarity between predicted and actual racial maps in 2010 than in the case of the $SNN_{1\to 2}$ model. Note that classification accuracy is a non-spatial index and two predictions with the

same values of classification accuracy can show different patterns. In the case of $SNN^{\star}_{1\to 2}$ inaccuracies are distributed among predominantly randomly located cells. Therefore, if we smooth predicted and actual maps the prediction accuracy of this model increases to 0.79.

Figure 1 shows the 2000-2010 change prediction obtained using the $SNN^{\star}_{1\to 2}$ model and its comparison to to the actual 2000-2010 change. The change is shown using two maps, one for 2000 and one for 2010; only cells that changed are shown. Pairs of maps having the original (300 m) and smoothed (3 km) resolution is shown. The overall character of change is best assessed from smoothed maps. Comparing predicted change (lower row in Figure 1) with the actual change (upper row) we observe a good agreement. The prediction overestimates the amount of change in the lower-left corner of the county and in few small areas across the county. However, overall, the $SNN^{\star}_{1\to 2}$ model yields a useful, actionable prediction.

4 Conclusions

The presented methodology forecasts change in racial geography of US cities a decade ahead. The method relies on transforming census population counts data into a categorical raster. In this new form the data resembles land use/cover data and its change can be foretasted using a method similar to those developed for land change forecasting. They are two differences between our method and those most used for land change, we don't use drivers (ancillary variables) and we handle non-stationarity differently. Our model uses only categorical racial maps data from two previous census years to predict a categorical racial map in the forthcoming census year. This is sufficient for this problem. In land change methods a non-stationarity is handled by using the projected abundance of land categories (this is what we do as well) followed by cell allocations. An algorithm assigns category to each cell in such a way as to optimize the global solution with respect to probabilities obtained from stationary forecasting. Greedy optimization algorithm is used to make a problem computationally tractable. We have tested this approach on racial data and obtained a forecast very similar to that yielded by our synthetic training dataset approach but at the higher computational cost.

An advantage of our method is its simplicity. The method is a straightforward temporal extrapolation from the past data in a manner that does not require any prior knowledge about the future (linear extrapolation is used to estimate abundances of categories in the predicted year). It also does not require a theory detailing the process that drives the change. This is a crucial advantage, because, currently, there is no viable and falsifiable theory of racial change on the local scale. Another advantage is a low barrier to using our method. The input data of racial classification into 11 categories for 1990, 2000, and 2010 and all conterminous US by county or by a metropolitan area is available for download from SocScape at http://socscape.edu.pl. This dataset will be extended to 2020 once block-level data from the 2020 Census become available. To make a prediction one needs to get the data for a region of interest (a region does not have to be a whole county or to be restricted to multiple neighboring counties) and to follow a procedure here. Although we used Mathematica for our calculations, two key functions, NetModel (for training the SNN model) and FindMaximum (for the constrained maximization) have equivalents in other computational packages. The main limitation of our model is the way it handles non-stationarity, and, in particular, the fact that we use linear extrapolation to estimate abundances of cells categories in the future. There already exist good estimation of abundance of different racial subpopulations in future years. One way to improve the current method would be to transform future estimations of

subpopulation abundances into estimation of population category abundances.

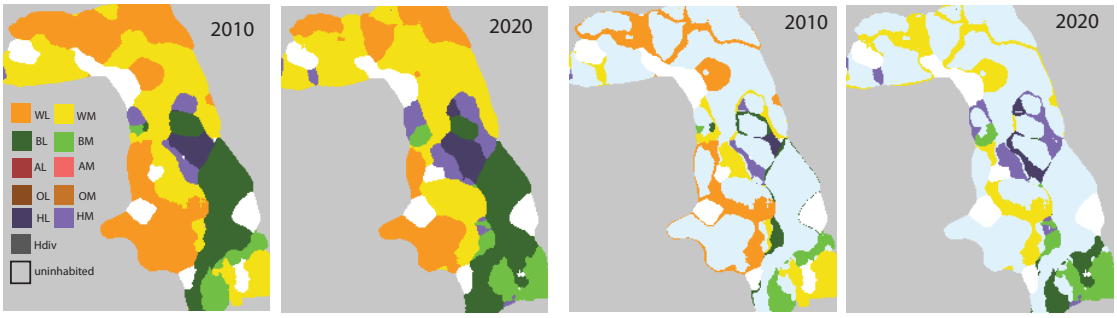

Figure 2 Predictions for the Cook County in 2020. (Left) Actual smoothed maps of racial maps in 2010 (actual data) and 2020 ($SNN^{\star}_{2\to 3}$ model prediction). (Right) Changes in values of cells' categories between 2010 and 2020; light blue color indicates cells that did not change their category between 2010 and 2020.

We also present our prediction for Cook County in 2020 (although 2020 is already in the past, the 2020 Census block-level data are still not available). For this we use the $SNN^{\star}_{2\to 3}$ model trained on the synthetic dataset (that uses 2000-2010 transition records and 2010-2020 transition frequencies) to predict racial distribution in 2020. Figure 2 shows the prediction. The change is shown using two maps, one for 2000 and one for 2010; only smoothed maps are shown. The pair of maps on the left shows racial distributions in Cook County in 2010 and 2020. The pair of maps on the right show the change between 2010 and 2020. It can be observed that the changes are located on the boundaries between population types. The dominant change is the expansion of WM area (yellow) at the expense of the WL area (orange). In the southern part of the county, the BM area (light green) replaces WM area (yellow). In the middle of the county we predict an expansion of Hispanic population in the western direction, the WM area is replaced by the HM areas, and, immediately to the east, the HN area is replaced by the HL area. Once 2020 Census block-level data are published this predictions can be checked against the actual data.

Finally, we note that our approach to spatially-explicit forecasting of racial change is very different from the recently published approach [3] that aims at addressing a similar problem. Georgati and Keßler paper also aims at spatially-explicit population forecasting including forecasting of future distribution of different migrants groups (in Copenhagen, Denmark), but attempts to achieve this by machine-learned (Convolutional Neural Network) regression model using a large set of ancillary data and direct, historical population counts. Once the two models mature it would be interesting to compare their corresponding advantages and disadvantages.

References

1 Laurie M Anderson, Joseph St Charles, Mindy T Fullilove, Susan C Scrimshaw, Jonathan E Fielding, Jacques Normand, Task Force on Community Preventive Services, et al. Providing affordable family housing and reducing residential segregation by income: a systematic review. *American journal of preventive medicine*, 24(3):47–67, 2003.

2 Anna Dmowska, Tomasz F. Stepinski, and P. Netzel. Comprehensive framework for visualizing and analyzing spatio-temporal dynamics of racial diversity in the entire United States. *PLoS ONE*, 12(3):e0174993, 2017.

187 **3** Marina Georgati and Carsten Keßler. Spatially Explicit Population Projections: The case of
188 Copenhagen, Denmark. *AGILE: GIScience Series*, 2:1–6, 2021.
189 **4** Mathew E Hauer. Population projections for us counties by age, sex, and race controlled to
190 shared socioeconomic pathway. *Scientific data*, 6(1):1–15, 2019.
191 **5** Steven R Holloway, Richard Wright, and Mark Ellis. The racially fragmented city? Neigh-
192 borhood racial segregation and diversity jointly considered. *The Professional Geographer*,
193 64:63–82, 2012.
194 **6** Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620,
195 1957.
196 **7** Carsten Keßler and Peter J Marcotullio. A Geosimulation for the Future Spatial Distribution
197 of the Global Population. In *AGILE 2017: 20th conference on geo-information science*, 2017.
198 **8** Günter Klambauer, Thomas Unterthiner, Andreas Mayr, and Sepp Hochreiter. Self-normalizing
199 neural networks. *arXiv preprint arXiv:1706.02515*, 2017.
200 **9** Ting Liu and Xiaojun Yang. Land change modeling: Status and challenges. *Monitoring and*
201 *Modeling of Global Changes: A Geomatics Perspective*, pages 3–16, 2015.
202 **10** Mathematica. Mathematica 12.0. Wolfram Research, 2020.
203 **11** Jacob J McKee, Amy N Rose, Edward A Bright, Timmy Huynh, and Budhendra L Bhaduri.
204 Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings*
205 *of the National Academy of Sciences*, 112(5):1344–1349, 2015.
206 **12** María Teresa Camacho Olmedo, Martin Paegelow, Jean-François Mas, and Francisco Escobar.
207 *Geomatic approaches for modeling land change scenarios*. Springer, 2018.
208 **13** Yanjiao Ren, Yihe Lü, Alexis Comber, Bojie Fu, Paul Harris, and Lianhai Wu. Spatially
209 explicit simulation of land use/land cover changes: Current coverage and future prospects.
210 *Earth-Science Reviews*, 190:398–415, 2019.
211 **14** J. S. Siegel and D. A. Swanson. *The Methods and Materials of Demography*. Emerald Group
212 Publishing Limited, 2004.
213 **15** Tomasz F Stepinski and Anna Dmowska. Spatially-explicit prediction of future racial distribu-
214 tion in major US cities: A machine learning approach. *submitted to Computers, Environment*
215 *and Urban Systems*, 2021.
216 **16** David A Swanson, GC Hough, Joseph A Rodriguez, and Chuck Clemans. K-12 enrollment
217 forecasting: merging methods and judgment. *ERS spectrum*, 16(4):24–31, 1998.
218 **17** Gerard C Wellman. Transportation apartheid: the role of transportation policy in societal
219 inequality. *Public Works Management & Policy*, 19(4):334–339, 2014.
220 **18** Hamidreza Zoraghein and Brian C O'Neill. US State-level Projections of the Spatial Distribution
221 of Population Consistent with Shared Socioeconomic Pathways. *Sustainability*, 12(8):3374,
222 2020.