The importance of historical linkages in shaping population density across space

L'importanza della dimensione temporale nella modellizzazione spaziale della densità di popolazione

Ilenia Epifani and Rosella Nicolini

Abstract This study aims to investigate the extent to which history matters in shaping population distribution across space. In the wake of the current literature, the idea is to model individual location preferences by focusing on selected local determinants (neighborhood, education, income, amenities and distance from the CBD), while also taking into account temporal and spatial dependence in location choices. Our preliminary results revel the importance of segregation factors in shaping recent population density distribution.

Abstract *Il presente studio analizza l'importanza della dimensione temporale come determinante della distribuzione spaziale della densità di popolazione. In linea con i contenuti della letteratura contemporanea, la nostra idea è di modellizzare le scelte di localizzazione degli individui in base a delle preferenze strettamente relazionate con alcune determinanti territoriali specifiche (tipo di quartiere, educazione, reddito, amenità presenti nel territorio e distanza dal centro d'interessi -definito come CBD-) in aggiunta a condizioni di autocorrelazione spaziale e di dipendenza temporale del processo di decisione. Quest'ultima, in particolare, garantisce la condizione di coerenza intertemporale delle decisioni prese dagli individui. I nostri primi risultati rivelano che la combinazione delle precedenti determinanti genera un effetto segregazione che assume un ruolo di primo piano per approssimare la distribuzione spaziale della densità di popolazione.*

Key words: Bayesian inference, Hierarchical dynamic model, Population distribution, Spatial random effects

1 Introduction and Data Description

Individual location choices are often driven by environmental factors beyond subjective preferences or budget constraints. As widely discussed in the critical contribution by [8] the characteristics of a neighborhood play an extremely important role in shaping individual choices. Neighborhood features matter both at present but also in throught reputation effects. For instance, think of a neighborhood that consolidates its reputation as a ghetto versus another that recently started hosting people of different ethnic origins. The former is more likely to impact individual location decisions compared to the latter. This study develops around the interplay between the relevance of environmental factors and historical dimensions in location choices

Ilenia Epifani
Politecnico di Milano, Dip. di Matematica, P.zza L. da Vinci, 32, I-20133 Milano e-mail: ilenia.epifani@polimi.it

Rosella Nicolini
Departament d'Economia Aplicada, Universitat Autònoma de Barcelona, Edifici B – Campus UAB, 08193 Bellaterra e-mail: rosella.nicolini@uab.cat

of the population in Massachusetts in a monocentric framework. The idea is to define a dynamic statistical model that allows for accounting for the way past values of some selected location determinants (referring to both a specific place and surrounding territories) may influence the present individual location choice. Thus, we are able to tract potential changes in the relative weight of local determinants and spillover effects in location-choice decisions across time. As argued in [1], [2], the geographical structure of Massachusetts simplifies the setting and, the historical and consolidated attractiveness of Boston as a Central Business District makes it possible to focus on a monocentric distribution function.

We built a Bayesian dynamic Log-Normal Model with random spatial normal frailties to investigate census tract data coming from the NHGIS project [6], the US census and the American Community Survey (ACS) for the 14 counties in Massachusetts for the period 1970-2010. The comparability of the geographical units is based on the information obtained from the TIGER/Lines Finder project. A census tract is identified as an area in which 1500 to 8000 habitants live (with an optimal size of 4000 persons). The tract is often split into two (or more) subtracts when goes over that limit or when the spatial territory is affected by other structural changes. Unfortunately, there is no a clear mechanism that allows for tracking with precision these changes over time, whereas their number increased progressively: census tracts were 1950 in 1970, 1193 in 1980, 1323 in 1990, 1361 in 2000 and 1472 in 2010.

Let $Y_{ij}^{(t)}$ be the population density (per square mile) of the jth census tract within the ith county of Massachusetts at decade t, for $i = 1,\dots,14$ and $t = 1970, 1980, 1990, 2000, 2010$. A preliminary explorative analysis of $Y_{ij}^{(t)}$ at county level reveals that the sample means per county exhibit temporal patterns different each from either county, but relatively stable within each county. Instead, the standard deviations proportionally shrink except for the most remote counties. These features suggest that the introduction of some spatial county random effects in the model can capture the heterogeneity among counties, whereas, the different patterns of the sample variances suggest a county-stratum variance heteroscedasticity.

In our analysis, the distance $D_{ij}^{(t)}$ of census tract i (in county j at decade t) from Boston is one of the key predictors of the density population and it is defined as a distance between two centroids: one being the centroid of the census tract with the highest population density (in a specific year) and the other being the centroid of any other remaining existing census-tract in the sample for that year. $D_{ij}^{(t)}$ is computed as the Euclidean value according to the geographical coordinates of the two centroids, as suggested by geographers. As our unit of reference (namely, census tracts) changes over time, distance $D_{ij}^{(t)}$ is a time-dependent predictor of the density population. Along with $D_{ij}^{(t)}$, we selected the ethnic composition, education and income as predictors for the population density. In particular, ethnic feature $M_{ij}^{(t)}$ of i th census tract is computed as the proportion of whites (over the total population). Income enters the statistical model via the income $I_{ij}^{(t)}$ per-capita per-census tract and via the Gini index $G_{ij}^{(t)}$ that measures the dispersion and inequality of income. Income $I_{ij}^{(t)}$ is not collected in 1970, so we can not use it in the regression for the first decade. Education in each census tract j enters the model via an index $E_{ij}^{(t)}$ obtained in the following way: first census-tracts were ranked according to the relative frequency of citizens having a high degree of education, then according to the relative frequency of persons having a low degree of education and hence the second value of ranking were substracted from the first. Finally, the presence of amenities is proxied by the proportion Z_i of free land (water areas) in county i: as discussed in [4], water can be considered a fundamental factor (or amenity) in creating recreational spaces for leisure time; moreover, as discussed in [2], it is reasonable to consider Z_i constant over time.

2 Bayesian Dynamic Hierarchical Log-Normal Models

Our empirical analysis is based on an augmented strategy. We begin with estimating a model without any kind of frailties. Then, we introduce some independent county-frailties. Finally, we elaborate a model with frailties having both a temporal and spatial dependence. In the light of previous remarks, we specified the following likelihood function

$$\begin{aligned}\log Y_{ij}^{(t)} | \boldsymbol{w}_t, \boldsymbol{\beta}_t, b_{0t}, b_{1,t}, \sigma_i^2 &\overset{\text{indep.}}{\sim} \\ &\overset{\text{indep.}}{\sim} \mathrm{N}\left(w_{it} + b_{0t} + b_{1t} Z_i + \beta_{1t} D_{ij}^{(t)} + \beta_{2t} M_{ij}^{(t)} + \beta_{3t} G_{ij}^{(t)} + \beta_{4t} E_{ij}^{(t)} + \beta_{5t} I_{ij}^{(t)} + \beta_{6t} I_{ij}^{(t)} \times D_{ij}^{(t)} \ , \ \sigma_i^2 \right), \quad (1)\end{aligned}$$

where $\boldsymbol{w}_t = (w_{1t}, \ldots, w_{14t})$ is the vector of 14 county random effects at time t. The $\boldsymbol{w}_t$s' capture the common features shared by the census tracts (in the same county), but they are different from county to county; each w_{it} summarizes all the determinants of the population density, either unobservable or observable but neglected, common to all the census tracts in the ith county.

In order to disentangle the random county effects from the evolution of the coefficients associated with the distance from Boston as unique determinant of population distribution, estimations have been performed in different steps:

(a) running the regression including only distance without frailties (i.e. assuming $\beta_{2t} = \ldots = \beta_{6t} = 0$ and $\boldsymbol{w}_t = \boldsymbol{0}$),
(b) running the regression with all selected covariates without frailties (i.e. assuming only $\boldsymbol{w}_t = \boldsymbol{0}$),
(c) running the regression including only distance with frailties (i.e. assuming $\beta_{2t} = \ldots = \beta_{6t} = 0$),
(d) running the complete regression with frailties.

In all of these specifications, a temporal dependence has been introduced at the fixed effect level, by assuming a random walk for the dynamic regression parameters $\boldsymbol{B}_t := (b_{0t}, b_{1t}, \boldsymbol{\beta}_{1t})$ as in [9] and, we get

$$\boldsymbol{B}_0 \sim \mathscr{N}(\boldsymbol{0}, 10^4 \times I), \quad \boldsymbol{B}_t | \boldsymbol{B}_{t-1}, \ \sigma_{b_0}^2, \ldots, \sigma_{\beta_6}^2 \ \sim \mathscr{N}(\boldsymbol{B}_{t-1}, diag\{\sigma_{b_0}^2, \sigma_{b_1}^2, \sigma_{\beta_1}^2, \ldots, \sigma_{\beta_6}^2\}), \ \forall t.$$

This prior is quite comparable with the canonical way to deal with the classical idea of adaptive expectations. Instead, we accomodate the heteroscedasticity in the counties by adopting the following multiplicative structure of the population log-density variances borrowed by [5]:

$$\sigma_{y;i}^2 = \sigma^2 \times \nu_i, \quad \sigma^2 \sim \text{Inverse Gamma}(0.001, 0.001), \quad \nu_1, \nu_2, \ldots, \nu_{14} \overset{iid}{\sim} \frac{\chi_r^2}{r}.$$

As for the county random frailties we test two alternative modelizations. To begin with, we assume a simple multiplicative hierarchical structure for $w_{1t}, \ldots, w_{14t}$ across both time and counties:

$$w_{it} \overset{i.i.d}{\sim} \mathrm{N}(0, \sigma_w^2), \quad \forall i, \ \forall t. \quad (2)$$

Alternatively, we model a spatio-temporal dependence between counties by including a space weight matrix A at a county-level to indirectly link the neighboring counties, via pseudo lagged Conditionally Autoregressive county random effects. The space weight matrix A is introduced as an adjacency matrix. Its cells are organized as follows: $A_{ij} = 1$ if i, j are neighbors, and $A_{ij} = 0$ otherwise. Then, A is used in the modeling of the county random effects:

$$w_{it} | \boldsymbol{w}_{t-1} \sim \mathrm{N}(w_{it-1} + \rho A \boldsymbol{w}_{t-1}, \sigma_{w;i}^2), \quad (3)$$

with initial conditions $w_{i0} \overset{iid}{\sim} \mathrm{N}(0, 10^6)\ \forall i$, and $\sigma^2_{w;i} = \lambda / n_i$ if i has $n_i \geq 1$ neighbors, $\sigma^2_{w;i} = \tilde{\sigma}^2_w$ if i is an isolated (or island) county. Our rational is that census-tracts belonging to a same county share some common features (e.g., urban regulation) while there may or may not be some dependence across counties in the way citizens form their expectations.

We complete the prior distribution for the remaining unknown parameters, adding a diffuse uniform prior distribution for σ_w on the range $(0, 10)$ in case of independent frailties (2). Instead, for dynamic-spatial frailties (3) we choose: $\tilde{\sigma}_w \sim Uniform(0, 10)$, $\lambda \sim$ Inverse Gamma$(0.5, 0.0005)$, $\rho \sim Uniform\,(1/l_{min}, 1/l_{max})$ where l_{min}, l_{max} are the minimum and maximum eigenvalues of the restricted spatial weight matrix A when excluding the two island-counties: Nantucket and Dukes. See [5] and the Manual of GeoBUGS, 2014.

3 Model Selection and Estimation

In order to assess the goodness-of-fit of our model, we computed the Bayesian p-values of the six alternative specifications on the base of the discrepancy measure by [3], for the density population on the logarithm scale: $\sum_{i,j,t} \left(log(Y^{(t)}_{i,j}) - \mathrm{E}(\log(Y^{(t)}_{i,j}) \,|\mathbf{Data}) \right)^2 / \mathrm{Var}(\log(Y^{(t)}_{i,j}) \,|\mathbf{Data}))$. Then, a Bayesian comparison of the six alternative specifications were performed by means of the DIC of each model and, for each model and each year, bu means of the percentage of the Bayesian census tracts outliers, i.e. of the census tracts whose actual population density falls into one of the two 2.5 percent tails of the marginal posterior predictive density. Overall, Bayesian-p value suggest that all our specifications perform relatively well by producing relatively good fits, whereas the DIC values reveal that the best empirical specification has to include ethnic composition, education, income along with the physical distance to Boston and the county random frailties. Both the "complete" models with county random frailties have quite similar DIC values, which are definitely better than the model with only distance and without frailties. Under this perspective, one can conclude that there is not a relevant difference in adopting a dynamic-spatial frailty-strategy rather than independent frailties. The Bayesian percentage outliers for each of the six models remains quite low (always under 5 percent) from 1980 to 2010. Instead, for the year 1970, results are at odds with the ones for the other decades.

For 1970, no covariate but distance to Boston seems to be able to model population density distribution. This result confirms that from 1970 to 1980 the land organization in Massachusetts passed through important structural changes, as anticipated by [2]. Figure 1 shows a clear tendency to identify the distance from Boston as the principal driver of population density distribution. Estimations referring to 1970 are a bit at odds with respect to the other decades. Instead, from 1980 onward, the results are consistent. The ethnic composition of the neighborhood and the level of income are two dominant and constant effects. High-income persons tend to settle in the area in which they can enjoy the possibility to rent or buy individual dwelling properties: these places are located far from Boston. In 2010 differences in education among the various census tracts also seem to become a statistically important discrimination factor in location choices, moving in the same direction as ethnic composition and income. County random frailties make the presence of natural amenities (here Z) ineffective. The interaction term between income and distance amplifies the attractiveness of each tract-unit, emphasizing the presence of a mass effect for a number of selected features (in this case, income). The idea is to capture the attractiveness of a destination focusing on some selected features that intervene in shaping individual preferences for location choices despite the physical distance from the CBD.

As far as the heteroscedasticity of the counties, we notice from the plots in Figure 2(a) that all of the econometric exercises provide consistent estimations of the clustered-per-counties heteroscedasticity of the census population density, with the highest variance in the most remote counties: Norfolk, Plymouth, Essex, Dukes, Suffolk, Middlesex and Barnstable are heteroscedastic with a low level of variance, Hampden,

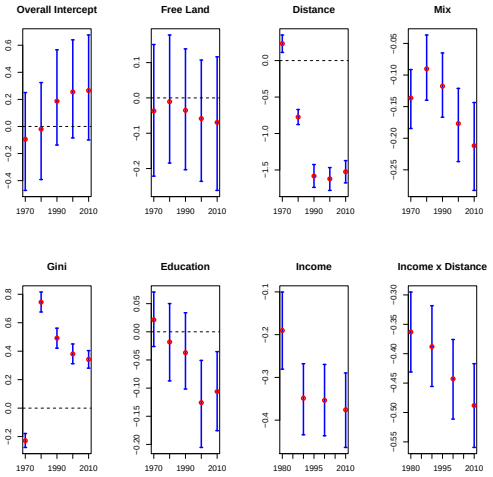


Fig. 1 Plot of the Bayesian estimates (posterior means) and posterior 90 percent Intervals of the regression coefficients in the *Complete Model with Independent Frailties* from 1970 to 2010. Data in 1970 for *Income* are not available. Legend: solid red circles: posterior mean.

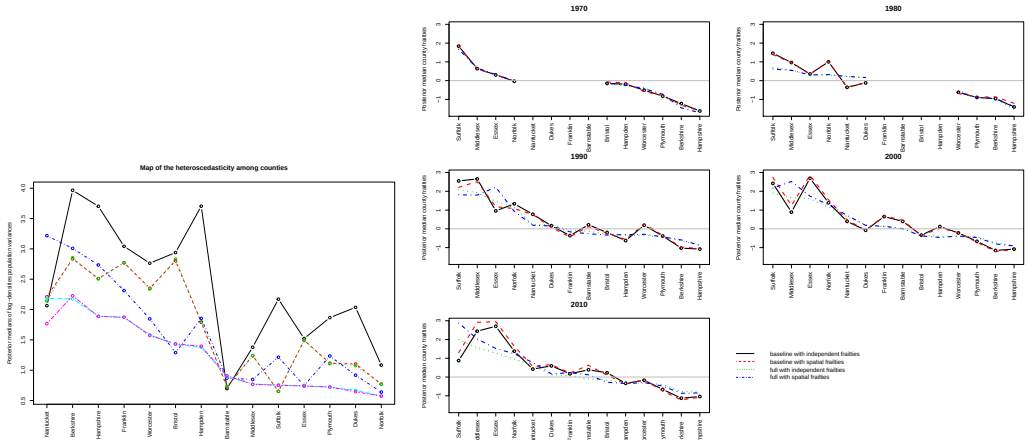
Bristol and Worcester have common variance, whereas Hampshire, Franklin, Berkshire and Nantucket are heteroscedastic with a high level of variance.

With regard to the county random heterogeneity, the picture of the posterior median values of the 14 county random frailties w_{it} in Figure 2(b) reveal that, on one hand the fitted frailties are quite identical under the two specifications, on the other hand, adding new local determinants of the density population along with the distance from Boston, compresses the frailty-predicted value toward zero. Although, under a temporal perspective, each frailty follows an autonomous evolutionary path, still, it is possible to detect some regularities. A structural break is definitely present between the first two decades (i.e. 1970 and 1980) and the rest (i.e., 1990, 2000 and 2010). Referring to the second period, the most remote counties (with respect to Suffolk) show frailty values that continue to be negative across time: the core counties (namely, those closest to Boston) display a strictly negative estimated median value, and the remaining countries –all far from Boston– either approach zero or are strictly positive. Hence, this study confirms that the populations of towns belonging to remote counties far from Boston consolidated low attractiveness in relation to Boston as the CBD across time.

4 Concluding Remarks

In this paper we propose a novel study to assess the determinants of population density distribution across space in Massachusetts. Our empirical strategy relies on the definition of a Bayesian framework in which we take into account temporal and county-spatial dependence for the assessment of population distribution at a census tract level over the period 1970-2010. Our main results confirm that the physical distance from Boston is the relative dominant covariate in defining population distribution across time. However, we detect an increasing segregation effect in modeling population distribution across time: the ethnic and income

characteristics of a neighborhood exhibit a relevant impact in the location choice of citizens. These discriminating features are strong enough to be robust and keep being relevant even in the presence of county random frailties aimed at capturing county features not directly embedded. Nevertheless, the county spatial dependence (built around the county frailty effects) does not improve the quality of the estimation if it is compared to the case of spatial-independent frailties. As for preliminary insights delivered from our results, one can conclude that the driving forces of population distribution are mostly associated with elements featuring segregation effects rather than spatial distance to a place of interest. Hence, potential actions aimed at smoothing population concentration (or congestion effects) should target the impact of segregation features rather than centering on accessibility issues. In policy terms, these new outcomes could deliver interesting suggestions for implementing policies aimed at reducing concentration or congestion effects as well as limiting the formation of spatial areas suffering from segregation effects.



(a) (b)

Fig. 2 Posterior Medians of the of the population log-density variances in Figure 2(a) and of the county random frailties in Figure 2(b) from 1970 to 2010. (Non available census-tracts in counties Barnstable, Franklin, Hampden and Norfolkfor for 1970, 1980). Counties are sorted horizontally in decreasing order of their posterior frailty medians under model (2) with all determinants.

References

1. Epifani, I., Nicolini, R.: On the density distribution across space: a probabilistic approach. J. Reg. Sci. **53**, 481-510 (2013)
2. Epifani, I., Nicolini, R.: Modelling population density over time: how spatial distance matters. Reg. Stud. **51**, 602-615 (2017)
3. Gelman, A., Carlin, J., Stern, H., Rubin, D.B.: Bayesian Data Analysis. CRC Press, Boca Raton, FL (1995)
4. Glaeser, E., Ward, B.A.: The causes and consequences of land use regulation: evidence from the Greater Boston. J. Urban Econ. **65**, 265-278 (2009)
5. LeSage J.P., Pace, K.R.: Introduction to Spatial Econometrics. CRC Press, Boca Raton, FL (2009)
6. Minnesota Population Center (2011). National Historical Geographic Information System: Version 2.0., Minneapolis, MN: University of Minnesota
7. Quigley, J.M.: Consumer Choice of Dwelling, Neighborhood and Public Services. Reg. Sci. Urban Econ. **5**, 41-63 (1985)
8. Topa, G., Zenou, Y.: Neighbourhood Effects versus Network Effects. In: Duranton, G., Henderson, J.V., Strange, W. (eds.) Handbook of Regional and Urban Economics **5**, pp. 561-624. Elsevier Publisher, Amsterdam (2015)
9. West, M., Harrison, P.J.: Bayesian Forecasting and Dynamic Models. 2nd edition, Springer, New York (1997)