Title
Discussion of “Measuring Housing Vitality from Multi-Source Big Data and Machine Learning”

Journal
Journal of the American Statistical Association, 117(539)

ISSN
0162-1459

Author
Banerjee, Sudipto

Publication Date
2022-07-03

DOI
10.1080/01621459.2022.2098135

Discussion of "Measuring Housing Vitality from Multi-source Big Data and Machine Learning"

Sudipto Banerjee

Department of Biostatistics,

University of California, Los Angeles, CA, USA.

February 8, 2023

I commend the authors on undertaking an elaborate analysis of a challenging data set. The article borrows information across multiple data sets to estimate the spatial-temporal distributions in "housing vitality", i.e., number of occupied houses, in Shanghai, China. The approach can be briefly described as comprising the following three steps: (i) use energy consumption data to label individual daily occupancy status; (ii) transform the pattern of occupancy status into monthly housing vitality indices that are subsequently aligned and merged with nightlight and land-use data; and (iii) spatially predict housing vitality across Shanghai. Probabilistic clustering of the energy consumption data is achieved using Gaussian mixture models in step (i) to identify homes that are occupied or not, while deterministic K-means clustering is used in step (ii) to estimate monthly housing vitality indices. Step (iii) involves dimension reduction of predictors using latent factors and idiosyncratic components, which are subsequently used to predict the number of occupied houses over time across Shanghai. The need for the dimension reduction appears to be the large number of predictors

consisting of Gaussian kernel-based spatial-temporal smoothed values of nightlight data and monthly effects, and their interactions with land-use variables.

The authors describe their methodology as "machine learning", a term that is nowadays used widely and somewhat indiscriminately to describe modeling and analysis of data using kernel machines. From the perspective of probabilistic machine learning, there are at least two issues of potential concern in the proposed approach. First, uncertainty propagation remains unaccounted for as we move through the three steps described above. Hence, uncertainty in the occupancy estimates obtained in steps (i) and (ii) are not accounted for in the predictive inference in step (iii). Second, I do not see exactly how spatial and temporal associations are being reckoned with in building the occupancy status of households from the daily energy consumption. A natural question, then, is how we can build a unifying modeling framework that can account for uncertainty in all of the above steps while also accounting for spatial-temporal associations at all resolutions. I outline some thoughts below.

Let $E_{it}(s)$ be a dynamic spatial process denoting the energy consumption for household i within a geographically referenced "location" s on day t. This location will depend upon the resolution at which the data is available and the type of inference being sought. For example, s could be a point-referenced location (e.g., latitude-longitude) of a building or it could be a pixel or region over which energy consumption has been aggregated or averaged. It may be important to accommodate the possibility of multiple households in a single location, especially in highly urban regions such as Pudong teeming with high-rise apartments, where a single location (even if it is the coordinate of a building) comprises several households (e.g., in apartments within a building). With this spatially explicit dependence, the authors' approach can be adapted to define a binary occupancy process,

$$Z_{it}(s) = \begin{cases} 1 & \text{if } E_{it}(s) \geq \tilde{E} \\ 0 & \text{if } E_{it}(s) < \tilde{E} \,, \end{cases} \tag{1}$$

where $Z_{it}(s) = 1$ indicates that household i in spatial location s was occupied on day t based upon whether the energy consumption exceeded a threshold $\tilde{E}$. Building upon ideas in Vanhatalo et al. (2021), one could consider mixtures of Gaussian process models to achieve spatially-temporally indexed clustering for energy consumption. This enriches the authors' approach by accounting for spatial-temporal associations while evincing reasonable values for $\tilde{E}$, which is used to label occupancy status as in the article under discussion.

The above idea still remains unsatisfactory in terms of accounting for uncertainty propagation through the subsequent modeling steps resulting in incorrect standard errors for predictive inference. In fact, it begs the question whether using a fixed threshold to label occupancy status is a desirable approach. An alternative, perhaps more traditional, statistical approach would be to resort to a survey-based sampling design to record occupancy status. However, designing such a survey may be too complex to obtain on a daily basis but may be feasible as a monthly aggregate. Regardless of whether the dependent variable is obtained from a survey-based sampling strategy or whether is has been computed using (1), let $Y_t(s)$ be the number of households with status "occupied" in time window t at spatial location s. A straightforward hierarchical model assuming a Poisson process specifies for each location,

$$Y_t(s) \overset{ind}{\sim} Poi(\lambda_t(s)) \; ; \quad \log \lambda_t(s) = x_t(s)^\top \beta_t + w_t(s) \; , \tag{2}$$

where $\lambda_t(s)$ is a spatially-temporally varying intensity function which accommodates space-time referenced predictors $x_t(s)$ and an underlying stochastic process to generate spatial-temporal dependence. The components of $x_t(s)$ can include extracted features (factors and idiosyncratic components) from smoothing kernels, fixed effects and their interactions with land-use variables as described in the article.

Bayesian inference will further introduce probability specifications for the possibly time-varying regression coefficients $\{\beta_t\}$ as well as other unknown parameters needed to specify the stochastic process for $w_t(s)$. Not all variables comprising $x_t(s)$ need to be space or

time-varying (e.g., certain land use variables are expected to remain static over the period of the study), but the time-varying β_t can help us evaluate the dynamic nature of the impact of these predictors on housing occupancy, while $w_t(s)$ will capture extraneous spatial-temporal variation attributable to unaccounted sources including variables not accounted for in the predictors. Bayesian inference using the spatial process model in (2) also offers fully model-based predictive inference at arbitrary spatial scales by sampling from the joint posterior $p(\{w_t(\mathcal{S})\}, \{\beta_t\}, \theta, Y_t(s^*) \,|\, y_{\text{obs}}) \propto p(\{w_t(\mathcal{S})\}, \{\beta_t\}, \theta \,|\, y_{\text{obs}}) \times p(Y_t(s^*)|\{w_t(s)\}, \{\beta_t\}, \theta)$, where s^* is an arbitrary unobserved spatial location, y_{obs} denotes the data on the outcome variable to condition on for inference, $\{w_t(\mathcal{S})\}$ denotes the realization of the spatial process over a finite set of spatial locations denoted by $\mathcal{S}$ and θ includes all parameters included in the prior specifications of $w_t(s)$ and β_t's. We first sample from the posterior distribution $p(\{w_t(\mathcal{S})\}, \{\beta_t\}, \theta \,|\, y_{\text{obs}})$ and, then, for each post-convergence value of these samples we sample the conditional predictive distribution $p(Y_t(s^*)|\{w_t(\mathcal{S})\}, \{\beta_t\}, \theta)$, which yields samples from the predictive distribution of $p(Y_t(s^*) \,|\, y_{\text{obs}})$.

Hierarchical modeling frameworks such as (2) can be adapted to a variety of inferential settings (see, e.g. Cressie & Wikle 2011, Banerjee et al. 2014). For example, the model applies to arbitrary spatial scales and the response $Y_t(s)$ can be treated as a count of occupied households at a coordinate (e.g., the number of households or apartments in a building) or over an overlaid grid to align with other predictors (e.g., brightness indices from nightlight data). The temporal scale, too, can be adjusted according to the scales where the predictors are available. Customary specifications for the dynamic spatial-temporal process can include auto-regression $w_t(s) = G_t(s)w_{t-1}(s) + \eta_t(s)$, where $\eta_t(s)$ are spatial processes independent over time (see, e.g., Stroud et al. 2001, Gelfand et al. 2005) for point-referenced locations. Dynamically evolving Markov random fields or conditional auto-regression models can be constructed should spatial dependence be modeled using adjacency matrices among neighbors (Waller et al. 1997, Ferreira et al. 2011).

Borrowing information from multiple sources to enrich inference, as has been undertaken by the authors' here, yields variables at disparate spatial-temporal scales. The authors encountered this issue with the nightlight data being available at a different resolution from the energy consumption data, while the land use variables may be available in yet again a different scale. This problem is referred to as spatial-temporal misalignment or the change of support problem (COSP) and a key advantage of process-based frameworks such as (2) is that uncertainty in aligning variables can be accounted for in the inference (see, e.g., Gelfand 2010, for a detailed review). While clustering methods such as K-means employed by the authors' can align variables, such methods do not allow the uncertainty to be quantified when the aligned variables are used in subsequent steps. Alternatively, if a joint spatial-temporal process for energy consumption $E_{it}(s)$ and nightlight data $X_{it}(s)$ is specified at every point-referenced location s, then model-based uncertainty quantification will be possible at arbitrary scales using either spatial interpolation or "kriging" (Stein 1999) or using stochastic sums or averaging over grid blocks as the case may be.

I conclude with some remarks about the specific analysis conducted in the article under discussion. Rather than using the energy threshold $\tilde{E}$ to identify occupancy status of households as a binary indicator, one could build a simple regression model with energy consumption as the dependent variable conditional on brightness indices from nightlight data as a predictor. While the aforementioned discussion on misaligned spatial data analysis can provide pointers to fully model-based inference, an even simpler approach will first interpolate the brightness indices from nightlight data to those locations in Pudong where energy consumption has been measured. We then fit the spatial regression model $[\text{Energy consumption} \sim \{\text{Nightlight brightness index}, \text{Spatial process}\}]$ to formulate a relationship between these variables. This relationship could then be used to predict energy consumption at all points in the Shanghai metropolitan area where nightlight data are available. This itself may offer a fairly robust interpolated map of energy consumption over

Shanghai including prediction intervals. Spatial locations that yield prediction intervals lying completely above an energy threshold will then be identified as occupied. Using the density of households at these locations one can form a vitality index. I would expect a strong association between household energy consumption and nightlight brightness, which may provide very robust estimates of household vitality index. Perhaps the index obtained in this manner will not be too different from the more complicated multi-step method proposed by the authors', especially from the standpoint of practical policy implications.

References

Banerjee, S., Carlin, B. P. & Gelfand, A. E. (2014), *Hierarchical Modeling and Analysis for Spatial Data*, CRC press.

Cressie, N. A. C. & Wikle, C. K. (2011), *Statistics for spatio-temporal data*, Wiley series in probability and statistics, Hoboken, N.J. Wiley.
URL: *http://opac.inria.fr/record=b1133266*

Ferreira, M. A. R., Holan, S. H. & Bertolde, A. I. (2011), 'Dynamic multiscale spatiotemporal models for gaussian areal data', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **73**(5), 663–688.
URL: *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2011.00774.x*

Gelfand, A. (2010), Misaligned spatial data: The change of support problem, *in* A. Gelfand, P. Diggle, M. Fuentes & P. Guttorp, eds, 'Handbook of Spatial Statistics', Boca Raton, FL: CRC Press, pp. 517–539.

Gelfand, A. E., Banerjee, S. & Gamerman, D. (2005), 'Spatial process modelling for univariate and multivariate dynamic spatial data', *Environmetrics* **16**, 465–479.

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, first edn, Springer.

Stroud, J. R., Müller, P. & Sansó, B. (2001), 'Dynamic models for spatiotemporal data', *Journal of the Royal Statistical Society: Series B* **63**, 673—-689.

Vanhatalo, J., Foster, S. D. & Hosack, G. R. (2021), 'Spatiotemporal clustering using gaussian processes embedded in a mixture model', *Environmetrics* **32**(7), e2681.
URL: *https://onlinelibrary.wiley.com/doi/abs/10.1002/env.2681*

Waller, L. A., Carlin, B. P., Xia, H. & Gelfand, A. E. (1997), 'Hierarchical spatio-temporal mapping of disease rates', *Journal of the American Statistical Association* **92**(438), 607–617.
URL: *https://doi.org/10.1080/01621459.1997.10474012*