Notes: Two Models of Multiscale Agglomerative Settlement

Phil Chodrow

May 22, 2017

1 Motivation

The purpose of these notes is to build on the insights of the paper Rybski et al. (2013) and Ema Strano that:

1. Human settlement displays local spatial correlations – sites are more likely to be settled the closer they are to existing settled sites.
2. Human settlements come in multiple "types," and the strength of the relevant local correlations may depend dramatically on the types of settlements involved.

We aim to specify probabilistic models of the dynamics of human settlement that display these features and allow us to learn system parameters from high-resolution data sets. In these notes, we develop two such models:

1. **Model 1** is essentially the same as Ema's original model in slide 11 of his Google Drive slides, modulo a few small structural tweaks. The model assumes the existence of precisely two settlement types, rural and urban. New settlements arise probabilistically at a rate depending on (a) their type, (b) their distance to other settlements of various types, and (c) a "background intensity" that accounts for settlement from sources outside the observation area. Each of these three factors have characteristic parameters that can be estimated from data. There is one additional parameter: a discrete size threshold T at which existing settlement clusters are considered "urban" rather than "rural." Parameter fitting proceeds via a simple EM algorithm in which the type of each new settlement is viewed as a latent variable, and T is selected via grid-search.
2. **Model 2** does not assume the existence of precisely two settlement types, and instead fits parameters to each cluster of existing settlement. It does this by viewing the settlement process as a spatially structured branching-like process, in which each new settlement is generated by an existing settlement in a distance-dependent way. The model assumes that each settlement has a spatial concentration parameter and an overall

importance factor, which is expressed as a mixing weight. Parameter fitting then proceeds via an EM algorithm similar Gaussian mixture modeling, in which each existing settlement is assigned a "responsibility" for each new settlement.

Relative to Model 1, Model 2 has the possibility to learn (rather than assume) the existence of multiple agglomeration scales, and the responsibility matrix may itself be interesting for modeling spatial linkages between sites. However, Model 2 may require more data to construct the responsibility matrix than Model 1, and could potentially lead to less interpretable results (such as if there is no discernible structured variation in the parameters for each settlement). My tentative recommendation is to try Model 1 first, since this will fit most neatly with Ema's existing theoretical framework and wouldn't require too much work to code up. If the data will support it, I expect that Model 2 could be quite interesting, and so I suggest that we plan to eventually try this one as well.

2 Model Summaries

Model 1

Model 1 most closely resembles Ema's original model, but with a few small tweaks that make its behavior a bit cleaner for both interpreting and learning the model parameters. It views growth as a two-stage process, in which each unsettled site is first assigned a potential settlement type (rural or urban), and is then settled with some probability depending on its type. So, the probability of an individual site being settled is:

$$\mathbb{P}(\text{labelled “rural”})\mathbb{P}(\text{rural settlement}) + \mathbb{P}(\text{labelled “urban”})\mathbb{P}(\text{urban settlement})$$

Of course, the types aren't observed, and we model them as latent variables.

The main question is how we should calculate $\mathbb{P}(\text{rural settlement})$ and $\mathbb{P}(\text{urban settlement})$; the label probabilities can be derived from these. I propose using a formula extremely similar to Ema's on Slide 11 of his Google Drive presentation. The main difference is that we pass the weighted sum of distances through a logistic sigmoid function to enforce normalization, which leads to more interpretable parameters, nice derivatives, and the ability to use standard tools from logistic regression. The probability of rural settlement in cell i is

$$\alpha_r \sigma \left(\sum_{j \in \mathcal{W}_r} d_{ij}^{-\gamma_r} \right) + \beta_r, \tag{1}$$

where σ is the logistic sigmoid function

$$\sigma(x) = \frac{1}{1 + e^{-x}}, \tag{2}$$

$\mathcal{W}_r$ is the set of all settled rural sites, d_{ij} is the distance between cells i and j, and where the parameters are:

- β_r is the "background rate" at which new rural settlements may appear without input from their nearby neighbors (for example, through migration).
- α_r reflects the importance of agglomeration effects compared to the background rate, and enforces an upper bound of $\alpha_r + \beta_r$ on the rate of rural settlement.
- γ_r reflects the spatial clustering of settlements – high γ_r corresponds to tight clustering.

As in Ema's original formulation, $\mathcal{W}_r$, the currently-settled rural sites, is taken to be the set of all settled clusters whose cluster sizes are beneath some threshold T. The probabilities of urban settlements are similar to (1), with urban parameters $\alpha_u, \beta_u, \gamma_u$.

Model 2

In Model 2, we instead view growth as a branching-type process. At each time step, each settled cluster may send out one or more "settlers" to locations governed by a probability distribution associated with that cluster. Then, we can model distance-weighted growth by stipulating that the probability that cluster C_j sends settlers to site i is

$$p_{ij} = \frac{1}{Z_j(\gamma_j)} \sum_{k \in C_j} d_{ik}^{-\gamma_j}, \tag{3}$$

where $Z_j(\gamma_j)$ is the partition function that ensures normalization. That is, the probability that cluster C_j sends settlers to site i decays with distance according to $d_{ij}^{-\gamma}$, similar to the Rybski paper.

In this model, each cluster gets two parameters: the distance decay exponent γ_j and a mixing coefficient π_j that tells how much absolute growth is driven by cluster C_j. For example, if the city has a large urban core with lots of growth packed tightly around it, we would expect that γ_j is large (tightly concentrated growth) and that π_j is large (lots of growth). On the other hand, if the city is highly disperse, with growth appearing in seemingly random places, we might expect that most clusters have γ_j small and that there's no cluster with very large π_j.

One cool aspect of Model 2 is that, if it works, it would allow us to estimate which existing clusters settled which new clusters, similar to the way that k-means clustering allows us to assign each point to a centroid. These estimates could themselves be pretty interesting, and may lead to some nifty network visualizations.

Another important point about this model is that it doesn't assume a hard divide between urban and rural growth regimes. Rather, the model could potentially discover two (or more!) distinctive growth regimes if, for example, the values of γ_j tended to cluster around two or more distinct values. This is probably the single main reason to be interested in this model – it can discover agglomeration scales rather than assume them.

3 Some Questions

1. How are the physics of these models related to Rybski's and Ema's original formulations? If the physics are substantially different then these models might not be useful,

so we should study some sample urban evolutions under these models before implementing learning algorithms.

2. Are these models too complex for the data? Model 1 has just seven free parameters, and so is likely tractable in cities that display decent amounts of growth over short time scales. Model 2 is potentially much more complex, and may therefore require more data. One way to reduce the number of parameters in Model 2 would be to assume that only clusters at some fixed size threshold can generate "settlers."

3. Model 2 has a fairly attractive physical interpretation in terms of settlement processes. Can we develop a similarly nice physical/economic interpretation for Model 1, or any similar model?

References

Rybski, D., Garcia Cantu Ros, A., and Kropp, J. P. (2013). Distance-weighted city growth. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 87(4):1–6.