A Stochastic Population Model Utilizing Regional Transportation Networks

David Meyer, Sam Rotter, David Shrekenhamer
Department of Mathematics, University of California, San Diego

Introduction

In the nineteenth century George Ravenstein empirically described the population distribution and migration trends in England (Ravenstein, 1885). He noted that when looking for work, workers tended to move randomly from place to place with preferences for the shortest paths, and he noted the existence of migration flows towards centers of industry. In the early twentieth century, Zipf posited the P1P2/D hypothesis (Zipf, 1946). The hypothesis states that given cities i and j the migration flow between i and j is proportional to the product of the populations and inversely proportional to the distance between the cities. This hypothesis leads to the development of a stochastic model to determine the population distribution of a geographic region. This type of model has come to be known as a gravity model (Smith, 1997), and uses the migration between two cities as transition probabilities in a Markov process. This gravity model is insufficient because it takes as parameters the very populations which it is trying to predict.

Modeling

In order to resolve the difficulties present in a standard gravity model we utilize the empirical observations made by Ravenstein to construct the transition probabilities associated with a Markov process. We define a matrix T with components

$$T_{ij} = \sum_{k=1}^{R} (W_{ijk} / D_{ijk}) \ , \ i?j \qquad (1)$$

and

$$T_{ii} = a_n J_i \ , \qquad (2)$$

where the a_n are three fittable parameters, R is the number of major roads from i to j, D_{ijk} is defined to be the distance along an existing road between adjacent regions i and j and to be infinite for regions which are not immediately adjacent along existing roads, W_{ijk} is the average width along the roads described in D_{ijk}, and J_i is a measure of job availability in region i. So, at each time step individuals move from region to region randomly with preferences for shortest paths and remain in a given region with a probability proportional to the number of jobs in that region. In this matrix we define preferences over remaining in regions at each time-step; this degree of freedom is omitted from many of the standard gravity-type models (Dorigo, 1983). We then define the normalized Markov matrix M where

$$M_{ij} = T_{ij} / \sum_{i} T_{ij} \ . \qquad (3)$$

So the ith column of M represents the complete set of actions an individual at region i may take at the next time step and the probability with which the individual will perform each action. Since M is a normalized Markov matrix we can determine the equilibrium probability that an individual will be in region i by examining the ith component of the normalized eigenvector associated with eigenvalue 1; this eigenvector defines the equilibrium population distribution for the network of regions being modeled (Doyle, 1988).

Goodness-of-Fit and Statistical Significance

Given the discrete nature of a population distribution and the dependence of the model on real data, non-linear fitting techniques are appropriate to determine the quality of the model and the statistical significance of the results. We apply a generalized, non-linear chi-square test to determine that goodness-of-fit and statistical significance of the model, whose predictions depend on the parameters $\{a_n\}$ and the data $\{Q_m\}$. The chi-square statistic is defined as

$$?^2 = \sum_i [(E_i - P_i)^2 / (\sum_j s_{ij}^2)] \qquad (4)$$

where E_i is the model-predicted population of region i, P_i is the observed population of region i, and s_{ij} is the standard deviation of changes in E_i due to errors in the measurements of the elements of $\{Q_m\}$, namely the road lengths and widths, the sizes of the job markets and the observed populations. The $s(\{Q_m: Q_m?P_i\})$ were determined numerically by varying these parameters within their respective confidence values, then finding the change in E_i for all i, and then taking the standard deviation of these changes in E_i. This test only applies when the changes in each E_i are normally distributed over the runs of the numerical simulation. The $s(\{P_i\})$ were determined by taking the standard deviation of the change in observed region populations over ten years of census data.

Empirical Results

We apply the model both to the counties of California (California Department of Finance, 2004) and to the major metropolitan areas of England (National Statistics, 2001). We divide the regions into three classes associated with the three fittable parameters. We form these classes from regions of similar geographic characteristics and economic status. Figure 1 shows the actual and predicted populations of California while Figure 2 shows the actual and predicted populations of England. The chi-square statistic described above is 32.3 for the cities in England and 13.9 for the counties in California. Both of these values indicate that we cannot reject our hypothesis.

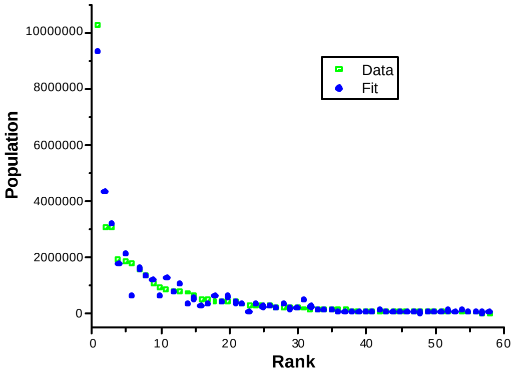

Fig.1 Plot of actual California populations in green and model predicted populations in blue

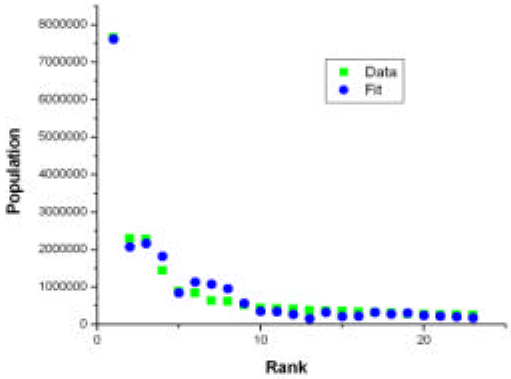

Fig.2 Plot of actual England populations in green and model predicted populations in blue

We can attribute a certain amount of error in the model to violations in the assumptions of the model. For instance, we assume that the major roads in an area represent the entire transportation network so we see large amounts of error between regions which have many small roads connecting them. Los Angeles and Orange counties demonstrate this type of deviation from observed data as can be seen in the first two data points of Figure 1 (ESRI, 2005). While a certain amount of error is to be expected we find that this model accurately describes both England and California despite large differences in their geographic location, economic status, and population dynamics. Given the differences between these two areas and the accuracy of the model, we conclude that our model demonstrates a strong correlation between population distribution and transportation networks.

References

Christaller, W. (1972) How I discovered the Theory of Central Places: A Report about the Origin of Central Places. P.W. and R.C. Mayfield, eds. Man Space and Environment. Oxford Univ. Press, 601-610

California Department of Finance: www.dof.ca.gov, copyright 2003 State of California

Dorigo, G. (1983) Push Pull Migration Laws, Annals of Association of American Geographers 72, 1-17.

Doyle, P. (1988) Random Walks and Electrical Networks, The Mathematical Association of America, Washington

ESRI website: www.esri.com, Copyright 1995-2005 ESRI

Karemera, D. (2000) A Gravity Model Analysis of International Migration to North America 32, 1745-1755, Routledge.

National Statistics website: www.statistics.gov.uk, Crown copyright material is reproduced with the permission of the Controller of HMSO

Ravenstein, E. G. (1885) The Laws of Migration, Journal of the Statistical Society 46, 167-235, England.

Smith, T (1997) Gravity-Type Interactive Markov Models, Journal of Regional Science, 37:653-708

Zipf, G. K. (1946) The P_1P_2/D hypothesis: on the Intercity Movement of Persons, American Sociological Review 11, 677-686.

Zipf, G. K. (1949) Human Behavior and the Principle of Least Effort, Cambridge, Addison-Wesley: Cambridge MA.