A HMM-Based Hierarchical Framework for Long-Term Population Projection of Small Areas

Bin Jiang[1], Huidong Jin[1,2], Nianjun Liu[1,2], Mike Quirk[3], and Ben Searle[3,4]

[1] The Australian National University, Canberra, Australia
bjiang@rsise.anu.edu.au
[2] NICTA, Canberra Lab, Locked Bag 8001, ACT 2601 Australia
Huidong.Jin@nicta.com.au, nianjunl@rsise.anu.edu.au
[3] ACT Planning and Land Authority
Mike.Quirk@act.gov.au, Ben.Searle@ga.gov.au
[4] Geoscience Australia

Abstract. Population Projection is the numerical outcome of a specific set of assumptions about future population changes. It is indispensable to the planning of sites as almost all successive planning activities such as the identification of land and housing supply, the release of land, the planning and construction of social and physical infrastructure are population related. This paper proposes a new hierarchical framework based on Hidden Markov Model (HMM), called HMM-Bin framework, for use in long-term population projection. Analyses of various existing suburbs indicate it outperforms traditional Cohort Component model and simple HMM in terms of less data dependency, output flexibility and long-term projection accuracy.

1 Introduction

Population Projection is the numerical outcome of a specific set of assumptions regarding future population changes [6]. The numerical outcome could be the total population of a specific area and sometimes can be further partitioned into different age and gender cohorts like 0-4 males and 5-9 females. The projection can be made either for 5 to 10 years or for a longer period, such as 20 or more years. The target area could be as large as a country or as small as a suburb.

Population Projection is indispensable because of the following points. Firstly, it plays a key role for government agencies to cost-effectively support development by the timely delivery of infrastructure, facilities and services. For example, in the development plan of a new suburb, the government has to determine whether it is necessary to release land to construct a new shopping centre and, if necessary, when, where and what size of land should be released then. Secondly, according to laws, population projection is compulsory, say, for land planning in ACT [1]. Thirdly, population projection results are also widely used by the private sectors, including retailers, property developers and investors. To predict population, we have to take account of a number of social, economic and political factors. It is almost impossible without certain assumptions [2].

M.A. Orgun and J. Thornton (Eds.): AI 2007, LNAI 4830, pp. 694–698, 2007.

A long-standing approach to conduct population projection is the so-called Cohort-Component Model (also called CC model hereafter) [6]. However, this model requires high quality data, which are often unavailable in many places especially for *small areas* like suburbs. Moreover, if a long-term population projection for twenty or more years is required, it is difficult to make appropriate assumptions, especially for the migration. In addition, this model only produces point values [1]. Such a kind of projection is not expressive enough, especially for decision-making support. Details about this model and our assumptions can be found in [2].

To overcome these limitations, we have developed a new HMM-based hierarchical framework in this paper. To mitigate data dependency such as migration rates, it describes population sizes as observations and takes other factors combined as a hidden variable. It estimates the complicated relationship between hidden states and observation from historical data. It also uses the population projections made by Australian Bureau of Statistics (ABS) for larger areas (LAs) to guide the projection for smaller areas. It can produce point values, as well as intervals or population probability distributions. We will illustrate our model on Canberra suburbs with a 20-year projection. Following ABS conventions, we call Canberra suburbs Statistical Local Areas (SLAs) hereafter.

A HMM is a joint statistical model for an ordered sequence of variables. It is the result of stochastically perturbing the variables in a Markov chain (the original variables are thus "hidden"). The Markov chain has discrete variables which indicate the "state" of the HMM at each step. The perturbed values can be continuous and are the "outputs" of the HMM. HMMs are commonly used in speech recognition [5]. A HMM includes the following elements: N, the number of hidden states in the model; M, the number of observation symbols corresponding to each state; A, the state transition probability distribution; B, the observation symbol probability distribution, so-called the observation probability distribution matrix; and π, the initial state distribution. More specific descriptions can be found in [5].

In the rest of this paper, we describe the proposed framework in Sec. 2 followed by the experiments in Sec. 3 and conclusions in Sec. 4.

2 The Proposed HMM-Based Hierarchical Framework

It is not trivial to applying HMM for population projection, as our preliminary results in Sect. 3 indicate a plain HMM performs quite bad. Actually, we need to overcome several problems including preparing appropriate observation variables for representation and training, the diversity of SLAs, the number of hidden states and constraints for projection. We propose a hierarchical framework based on HMM (termed HMM-Bin hereafter) to overcome these problems. As illustrated in Fig. 1, HMM-Bin has different HMMs for projection for different kinds of SLAs, automatically grouped by clustering. It uses population

[1] a series of specific values.

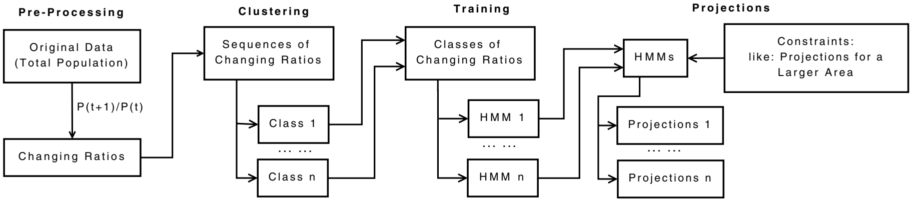

Fig. 1. A high-level view of our proposed HMM-Bin framework

projections for a larger area to guide projection for smaller areas. We describe how this framework overcomes these problems as below.

Pre-processing. The normally available data is total population by age and gender for a target area or a set of SLAs. However, because different SLA has different geographical size and hence can support different size of people, we use $r_{t+1} = \frac{P(t+1)}{P(t)}$, where $P(t)$ is the total population value at year t instead to avoid the absolute total population size and uncover something in common: *degrees of change.*

Clustering. Given the changing ratios, the degrees of change, some SLAs' population increases, some SLAs' population decreases while some SLAs' population fluctuates greatly. It is acceptable to take all the SLAs' changing ratio sequences as a whole class to train a plain HMM, but obviously it is not natural because differences of population change pattern between different SLAs do exist. So our solution is clustering these changing ratio sequences using clustering method like K-Means clustering algorithm with Euclidean Distance [4]. We have also experimented *WITHOUT* this step and the results (see Sec. 3) show that this step is quite necessary.

Training. After grouping SLAs into different classes, their changing ratios will also be grouped into different classes and each class has a class centroid sequence with the same length. According to the quantity of sequences within each class, we decide whether or not to build a HMM for that class. If the training data is insufficient, we will skip that class since HMM training needs adequate training data. We use Baum-Welch method for training a HMM. To use this method, we have to specify two things first as below.

Type of Observations: Since Cohort-Component method produces very specific point values of population size and its results have no flexibility at all. Therefore, it is necessary to produce more flexible results like intervals or population probability distribution. Provided that any probability distribution could be effectively approximated by a mixture of Gaussian distributions [3], it is arguable to adopt mixtures of Gaussians as our HMM observation type.

Number of Hidden States: The number of hidden states can be determined by examining the histogram of the changing ratios of different classes. Our approach is: (1) visualise the changing ratios with histograms and have an initial guess of how many Gaussian distributions needed to cover them; (2)

then do the training to see, within these initial guess values, which one will generate higher likelihood and less overlap among Gaussian distributions.

Projection. After the training step, the classes having sufficient training data will get their Hidden Markov Models with trained parameter sets. Here are the projection steps: (1) Fit out the relationship between the LA's historical changing ratio sequence and class centroid sequences: centroid sequence = f(LA's historical changing ratio sequence); (2) Get the future class centroid sequences using LA's future changing ratios and the fitted function f; (3) Use the projected centroid sequences as observations to guess the future hidden state sequences for different classes of SLAs in the target future period; (4) Use the projected hidden state sequences, we can get the corresponding future observations, which are then formed to be intervals of 95% confidence level. Since what we have projected are changing ratios, in order to get population sizes, we have to multiply the changing ratios with previous years' population sizes.

3 Experiments and Discussion

We examine HMM-Bin on existing Canberra suburbs and compare it with a plain HMM and the CC model. Our experiments are based on ABS' data for ACT SLAs' historic total population sizes from 1986 to 2000, ACT historic total population sizes from 1994 to 2000 and ACT population projections 2001 to 2020. Our target period is 20 years from 2001 to 2020. We use K-Means clustering algorithm with Euclidean Distance and Kevin Murphy's HMM MATLAB toolbox. The larger area (LA) discussed in the projection step is ACT in this case and its population projections made by ABS are used as constraints.

We only conduct projections for two classes [2](Class 1 and Class 2) that have enough training data. The ABS population estimations for these SLAs for 2001 to 2004 have been used as ground-true values. There are 92.50% out of Class 1 and 95.65% out of Class 2 SLAs have been covered by HMM-Bin projections at the confidence level of 95% while the coverage percentage for NON-Clustering is just 52.05%. Besides the coverage comparison, we also compare the relative difference between the mean sequences generated by HMM-Bin and the ground-true values. If we take 4% relative difference as a threshold, Class 1 and Class 2 have over 70% SLAs within this threshold while there is no SLA from the all-in-one Class within this threshold. The majority ($\geq$ 80%) of Class 1 and Class2 SLAs have relative differences within 6% while that of all-in-one Class have relative differences within 10%. These results show the weakness of a plain HMM directly applied *WITHOUT* the clustering step.

The comparison with CC model also has been done to several SLAs and details can be found in [2]. In the case of long-term projection, HMM-Bin projections are smoother than CC results which change more rapidly. HMM-Bin only produces 10% to 13% relative differences while CC produces over 40% relative difference

[2] After clustering, there are 6 classes of SLAs but only 2 of them have enough training data.

compared with the mean sequences. The reason why CC changes rapidly is because it is so hard to make assumptions about the future population changes especially for the migration part, for which it is almost impossible to predict accurately.

4 Conclusion

In this paper, we have proposed a HMM-based hierarchical framework for long-term population projection. We have evaluated it on various Canberra suburbs and compared with the traditional CC Model and a plain HMM. The HMM-Bin framework could generate more accurate population projections comparing some ground-true data from ABS. Moreover, the HMM-Bin framework has low dependency of data availability. This is quite useful for those *small areas* lacking of high quality of census data like suburbs in Australia. Furthermore, HMM-Bin framework produces flexible outputs in the form of intervals (population probability distribution in general), instead of specific point values generated by other models like the traditional CC model. Last but not least, because HMM-Bin framework has less requirements of data, it is more suitable than CC model for long-term projections for small areas like suburbs. Actually, it is quite hard to make assumptions especially for migration for CC model and these assumptions normally have low accuracy. Future effort may be put in enabling HMM-Bin to handle disaggregate population components.

Acknowledgment

We are grateful to Mr. Phil Jorritsma from ACT-PLA [3] for help in domain knowledge. We also appreciate Australian Bureau of Statistics for free data and financial support from ACT-PLA and NICTA, which is funded by the Australian Government's Backing Australia's Ability initiative, in part through the Australian Research Council.

References

1. ACT Parliamentry Counsel. Land (Planning and Environment) Act (2007)
2. Jiang, B.: Better Long-term Population Projection. Master's thesis, the Australian National University (June 2007), Available at `http://feng.marco.jiang.googlepages.com`
3. Jin, H.-D., Leung, K.-S., Wong, M.-L., Xu, Z.-B.: Scalable model-based cluster analysis using clustering features. Pattern Recognition 38(5), 637–649 (2005)
4. MacQueen, J.B.: Some Methods for classification and Analysis of Multivariate Observations. In: Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297 (1767)
5. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE 77(2), 257–286 (1989)
6. Smith, S.K., Tayman, J., Swanson, D.A.: State and Local Population Projections: Methodology and Analysis. Kluwer Academic Plenum Publishers, Boston, MA (2001)

[3] ACT Planning and Land Authority.