Mining Satellite Images for Census Data Collection: A Study Using the Google Static Maps Service

Frans Coenen
Department of Computer Science, The University of Liverpool, Liverpool, L693bX, UK

1. INTRODUCTION

A census is a mechanism for acquiring and collecting information about a population. It is a mechanisms that is widely used with respect to a variety of national and local government, management and planning activities. The most important element of a census is population count. However, Census collection and the associated post processing of the retrieved data is expensive. The UK Office for National Statistics (UKONS) reports that the UK 2011 census cost some 480 million [1]. The US 2010 census is reported to have cost $13 billion [2]. The cost of census collection is also increasing; according to the Australian Bureau of Statistics the Australian 2006 census cost around AUD 300 million; whilst the 2011 census cost around AUD 440 million. Cost with respect to rural areas is typically greater than in urban areas because the communication and transport infrastructure in rural areas tends to be less well developed. There is also often a lack of good will on behalf of a population to participate in a census, even if they are legally required to do so, because people are often suspicious of the motivation behind censuses.

Census collection is therefor challenging. One solution is the usage of technology, namely the internet. However, many people remain unconnected to the internet. In the context of the UK 2011 census it was found that the most frequently cited reason for households not to have internet access was because of a ”life style” decision not to do so. In less affluent parts of the world internet accessibility and usage is much lower (although arguably set to increase). Internet based census collection requires those completing the questionnaires to be literate, not necessarily always the case. An alternative solution is population estimation. Population estimation has been a subject of researched amongst the Geographic Information Systems (GIS) and remote sensing communities for some time. Estimation can be done using areal interpolation where existing census information concerning some geographic area is used as input to an interpolation algorithm to obtain a population estimation for a wider or alternative geographic area. Here we define an area for which we know the population size according to some set of relevant attributes and then (say) perform linear regression to produce a model that we can then used with respect to other areas that subscribe to the same attribute set. This seem like a good idea, however, the question remains as to what this attribute set should be comprised off, we need an attribute set based on information that is readily available at low cost.

The solution proposed here [3] is to build a classification model that can be used to predict household size using satellite image data. Although this will not work in cities where it s difficult to distinguish buildings in terms of number of inhabitants, it will work well in rural areas (where census data collection tends to be more of a challenge). To test this idea training data was acquired from a rural area of Ethiopia. The training data featured the locations of households where household size was known, therefore relevant satellite images could be obtained. Google Earth does not readily facilitate the automated extraction of satellite imagery, instead we used the Google Static Map Service. This features an API that allows users to download satellite images (one image at a time) specified according to various parameter settings, namely: latitude and longitude of the centre of area of interest, image size (in pixels) and the zoom Level (level of detail). We used an image size of 12801280 pixels (kLevelOfDecomp) and a zoom level of 18. Each household was then surrounded with a 256256 pixel bounding box defined so as to cover average household (by superimposing a box we do not have issues with irregular shaped household plots). In this manner a set of household images was produced.

A variety of mechanism were considered with which to represent households: frequent sub-graphs extracted from quad tree representations, colour histograms and Local Binary Patterns (LBPs). In the first case a set of quad trees (one per household image) was first generated to which Frequent Sub-Graph (FSG) mining was applied using a support threshold . The frequent sub-graphs were then used to generate feature vectors from which predictors could be generated. Using the Colour Histograms technique seven different histograms (red green blue, hue, saturation, value, grayscale), 32 bins per histogram, were generated and concatenate them together to give 224 attribute feature vectors (327=224), one per household. LBPs with eight neighbours and a radius of 1 were used. The conducted evaluation indicated that the LBP representation produced best results.

Once a prediction model had been built it could be applied to a much wider area. A village, and its surrounding lands in the Ethiopian hinterland, was selected. In 2011 it was reported that this village comprised 459 households and a population of 3,223 (thus ground truth data was available). 600 Satellite images were collected covering the area using the Google Static Map Service API. Note that the satellite mage data was from 2013, two years after the census, and collected using the Google Static Map Service API (image size of 12801280 pixels and zoom level = 18). The images were downloaded in an iterative manner, image by image, using a 320 pixel overlap designed so that every household will appear in its entirety in at least one image. Thus each downloaded satellite images could contain zero, one or more households.

The segmentation was conducted using a number of image masks. After segmentation a set of households images was obtained, each identified by a central latitude and longitude surrounded by a ww box (w=256, same value as used for classification model training). Boxes will be smaller and/or non-symmetrical near the edges of each image. Knowledge of the central latitude and longitude, and box size, was used to remove duplicate household images. A total of 526 households were detected including duplicates. Duplicate de-

tection identified 100 duplicate households, thus 426 out of a ”known” number of 459 households were identified. Suggested reasons for the discrepancy were as follows. The two-year time difference between the ”ground truth” survey and satellite images; a period during which some households may have fallen into disuse (manual inspection of a proportion of the collected satellite images indicated that some households did indeed appear to be roofless thus supporting this conjecture). Inspection of the satellite imagery indicated that a small number of buildings were very poorly defined and in some cases not segmented correctly. It was also possible that the duplicate household detection mechanism had detected some duplicates that were in fact not duplicates (although no evidence for this was found).

The prediction model generated earlier was then applied and a population size of 2760 predicted (compared to a ground truth of 3223). The reason for the discrepancy in this case might be because the data from which the prediction models were generated might not reflect the data to which they were applied as closely as anticipated. The two-year time gap between the date of the census collection (2011) and the date of the satellite image extraction (2013) might also have had an effect. As noted above, manual inspection of a number of images showed signs of derelict (abandoned) households. It may thus be the case that between 2011 and 2013 depopulation had taken place and that the produced population estimates were in fact a better reflection of population size than initially thought (there have been recent reports on depopulation in rural Ethiopia). Whatever the case, although the population estimations produced were not as accurate as the ground truth census data (this was to be expected), the proposed method offered significant cost and time-savings.

1.1 Acknowledgements

I would like to thank Kwankamon Dittakan of the Faculty of Technology and Environment at Prince of Songkla University (PSU) in Thailand for valuable support with respect to the work presented in this short paper.

References

[1] Office for National Statistics. National population projections, 2010-based statistical bulletin. Technical report, Office for National Statistics, 2011

[2] L.B. Shestha and E.J. Heisler. The changing demographic profile of the united states. Technical report, Congressional Research Service, March 2011.

[3] 3. Dittakan, K., Coenen, F., Christley R. and Wardeh, M. (2013). Population Estimation Mining Using Satellite Imagery. Proc. Int. Conf. Data Warehousing and Knowledge Discovery (DaWaK’13), Springer LNCS 8057, pp285-326.