Model for Small Area Estimation

Yacoub Nuseibeh[1,2], Raneen Herzalla[1,3], and Mohamed Almarri[1,4]
[1]*Insights and Foresights, Statistics Centre - Abu Dhabi, UAE*
[2]*E-mail: yhnuseibeh@scad.gov.ae*
[3]*E-mail: rhherzalla@scad.gov.ae*
[4]*E-mail: maalmerri@scad.gov.ae*

Summary

The number of individuals living in a given country is essential for decision makers to study, plan and implement policies and initiatives. Usually, this number is produced by the census once every 10 years. To better handle timeliness and accuracy of population estimates, the Small Area Estimation model was created. The standard population models like the cohort growth method does not work well for the case of Abu Dhabi or any other state where there is a majority of expatriates residing in and migrating out. This model is primarily based on utility consumption data. This makes possible the ability to obtain estimates on a monthly basis. However, the input data has biases, for which the correction factors were applied in order to mitigate. The final estimates obtained for the residential population are composed of the total population, distributed by district along with utilizing big data sources for demographics.

Key words: population estimation; small area; admin data; water consumption; data science

1 Introduction

The most precise method to obtain population size and characteristics is through the census as a baseline. Due to this method being costly and time-consuming, the Statistics Center at Abu Dhabi (SCAD) created a new estimation approach to provide policymakers and planners with up-to-date population indicators.

The absence and irregularity of receiving significant admin data – such as identity data – limits the ability of SCAD to use classical methods to estimate population. The standard cohort method is only applicable for the UAE nationals; however, expatriates' population is a majority of the population in the UAE including Abu Dhabi. These residents' backgrounds and characteristics are continuously changing due to several factors such as economic conditions, geopolitical issues, and government policies. Such unusual conditions lead to the use of assumptions and estimation models that could not represent the accurate population estimate and characteristics for Abu Dhabi nor its distribution in a useful way for decision makers.

In 2021 SCAD launched the Insights and Foresights Platform (IFP) which was built by an experienced team made of Data Science & Application Development teams. The IFP team built many machine learning and Hybrid models ranging from economic forecasting to what if scenarios for population growth. Typically,

SCAD works on modernizing statistical production models by utilizing alternative available admin data and unlocking big data capabilities to support the classical model (usually used in NSO's) to estimate the total population numbers in Abu Dhabi, as well as allocating them properly on a geographic level.

The new approach is: "Small Area Estimation (SAE) model" which aims to calculate the total population estimates by district in the Abu Dhabi emirate using utility consumption data as a base, and other admin sources to adjust for demographics. Non-Emirati residents will be estimated depending on the total population and Emirati resident estimates. Abu Dhabi emirate is characterized by a majority of non-Emirati resident population. Therefore, in-and-out migration plays a critical role in determining the patterns captured in Population indicators. Moreover, Abu Dhabi Emirate is progressing at a rapid pace where many areas become urbanized within a short time frame and that changes the structure of the population allocation in the emirate. This imposes a challenge when relying on census data which can be up to 10 years old as a baseline for the current population estimations.

2 Model Approach

2.1 Data Collection and Processing

Data was collected from different sources and combined to build the base of the model. This included sources such as the population database, Labor Force Survey and Household Income & Expenditure Survey. We also considered different admin registers, such as water consumption, Identity data, health insurance, and schools. Other sources such as telecom data was not available to the team.

2.2 Model Approach

The overall model approach is to use whatever good quality data available and then fill the gaps with data science and wrap the whole model with data science in order to create a dynamic and agile model. The first step in the model is to estimate the total population using water consumption. This large piece (the total) is then broken down into chunks which are allocated, adjusted, validated and enriched based on the admin data available. The enrichment part covers demographics of the population. This approach can be illustrated as below

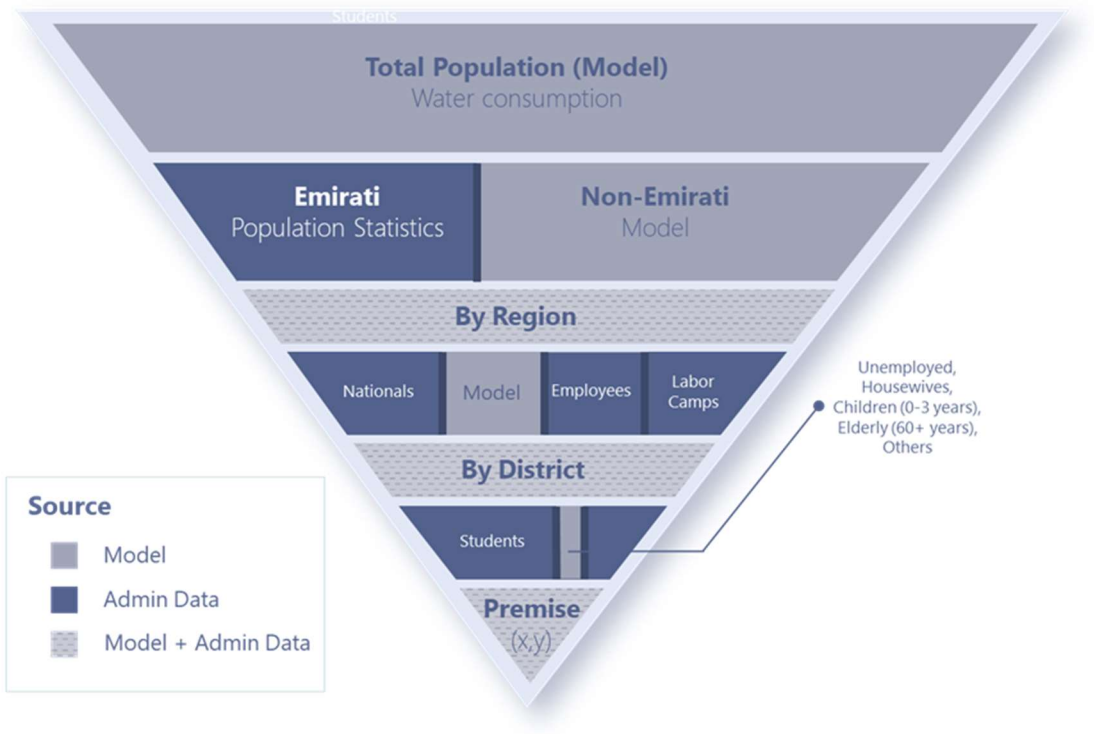

Figure 1. Model Approach. Classification Attributes: Residents status (UAE National vs. Non-UAE National status) Gender, Nationality, Region, District, Premise Type

2.3 Calculations and Adjustments

To calculate the Total Population, one starts with the monthly water consumption on a premise level, filtered for the residential sector. After that, it gets treated for potential outliers (zeros and excessive consumption), missing data, and duplicates. The next step is to convert from consumption amount to number of people. In order to do this and estimate the population, the total consumption and the average consumption per person should be known. From these two variables, the population can be calculated as follows:

$$total\ population = \frac{total\ utility\ consumption}{average\ consumption\ per\ person}$$

Since the utility consumption is on a monthly basis, the question of seasonality arises. By plotting the monthly water consumption for each year, we observed some fluctuations across the seasons and over the years. To account for this, a 12-month moving average was applied to the water consumption which allows to correct anomalies (extreme peaks and low points) and remove the seasonal variability (which showed an increased consumption in summer).

For the average consumption per person, as the data was not available, it can be calculated as follows:

$$average\ consumption\ \text{per}\ person = \frac{average\ consumption\ per\ household}{average\ household\ size\ (family\ members)}$$

Where the average household size was taken from a Household Income and Expenditure Survey (HIES) dataset that provides them by type of premise (villa, apartment, etc.) and resident type (Emirati, Non-Emirati). Given this split, the average consumption per household was also aggregated by premise type and resident (this was done after observing a difference in average consumption based on these 2 factors). As of this point, the total *residential* population is calculated for each premise, which can then be aggregated to district.

The penultimate step is to correct the resident distribution, since we know that for water consumption, some rented properties are registered under an Emirati-resident contract. Data was obtained on the distribution of the property owners and corresponding split between Emirati and Non-Emirati residents. Thus, this ratio was used as the correction factor to adjust the resident proportion and correct the bias that the input data source had. Finally, the labor camp workers are added to the estimate to move from *residential* to *total* population. Total labor camp estimates by region (which is one level higher than district) was used in this case, along with another source that provided each labor camp by capacity. These region estimates were then pro-rated to each camp based on its capacity. Now the total population has been estimated by resident, down to the Small Area (district).

2.4 Validation

To validate the estimates, they were compared with different models and admin sources. Given that one admin source can only cover a proportion of the population at a time, this was considered when doing the comparison. The variance calculated was below 10%. Additionally, randomly selected small samples were tested, and the accuracy of the results were around 62%. Enhancements and refinements are being implemented to improve the outcomes yet for the total population figures the accuracy was much higher (97%). This was validated versus National Identification Data which has aggregated totals but no distribution within the smaller areas. These results triggered the statistics population team to change their current model.

2.5 Enrichment and Further Adjustment

Bringing demographics into the picture, one can further enrich and adjust the estimates data for resident, nationality and gender using a wide variety of admin data sources. Overall, data on Emirati residents and their demographics from the official population statistics was incorporated into the model, given the nature of their stability as a population count. For residents, rules were applied on a premise level based on household type and estimated household size (these rules were based on the generally known living arrangements/standards of each). For gender, data merged from national identity and education (for students) were used to extract the gender distribution per district. Firstly, the gender distribution by district is calculated for both Emiratis and Non-Emiratis. Secondly, the percentage of female population is applied per district. Lastly, the male population is the total population estimate minus the female population – this would be done prior to including labor camp estimates. For nationality, students' data can be used to generate nationality distributions per district, and used to replace the distributions provided by the water

consumption data (only for accounts classified as married) and then merged with the rest of the accounts. For labor camp workers, the resident distribution was derived as 100% non-Emirati for resident status and 96% male / 4% female for gender. Finally, a small sample was taken and validated against these demographics. The labor camp data is based on several admin and labor force surveys estimates (LFS).

3 Output

Since water consumption data is available on a monthly basis, the estimates can also be automated and output monthly. To visualize the population and its distribution, the data on a premise level can be imported into a visualization tool and made interactive. Each dot represents a household. An estimated figure of number of people, their residency status, their gender, & their age will be displayed when chosen.

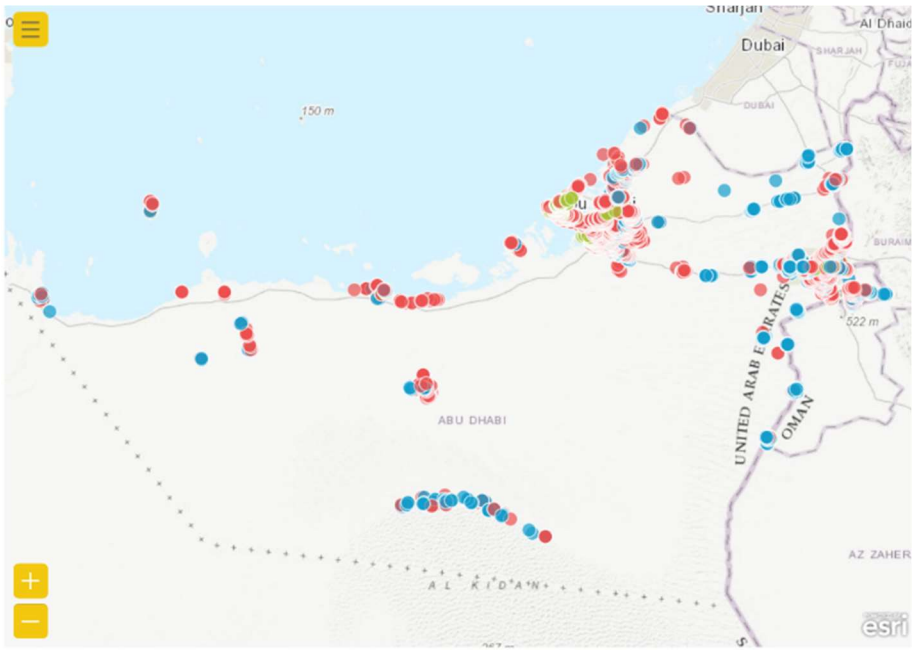

Figure 2. Population estimates plotted on a map, with interactive features to drill up and drill down (data tables left out due to confidentiality)

4 Concluding Remarks

In the Small Area Estimation, a model was created to estimate the population based on the utility consumption. This allows to obtain the population estimates on a monthly basis instead of the census data which is normally conducted every 10 years. However, the input data had biases and the correction factors were applied in order to mitigate them. The final estimates are: The total population (with demographics), Emirati Nationals, and Non-Emirati Nationals as

$$NonEmirati\ Nationals\ =\ Total\ population\ -\ Emirati\ Nationals$$

Furthermore, historical water consumption data as a time series can lead to future projections of the population.

The IFP team in SCAD is pushing the boundaries of traditional statistics, by blending data science and traditional methods or theories. This has proven to be a very effective way as most models built such as this one covered essential needs for decision makers who are increasingly relying on the most updated and accurate figures.