Contribution ID: **29** Type: **not specified**

Synthetic populations for personalized policy

Tuesday, 21 September 2021 16:50 (30 minutes)

Public policy design generally targets ideal households and individuals representing average figures of the population. However, statistics only make sense when referring to large numbers, less so when we are trying to represent real people belonging to the actual population. In fact, referring to the characteristics of the average citizen, the policy maker loses the capacity to represent the diversity of the population at large, negatively affecting minorities and under-represented people.
Statistics over the population are usually given as univariate figures. Typically, knowing that e.g. in a certain area live 55% women and 30% university educated people do not give a high quality information for the distribution and we may actually misinterpret what the real issue is.
One way to improve the representation of the diversity is to recur to multivariate distributions in spatial modelling, e.g. creating high quality aggregates for specific use cases.
Using real data to give these representations poses important privacy concerns, because knowing the combination of features in certain areas might give away the identity of some citizens.
In recent years, the performances of supercomputers skyrocketed, and at the same time the access for data scientists to high performance computing technologies has been democratized, offering to policy makers the unprecedented opportunity for creating tailored policy using a completely synthetic population.
Policy simulation models can take as input synthetic individuals that resemble the actual ones but are stripped out of their identities, as they are synthetic by design. Synthetic individuals are created inter-linking census data, behavioural surveys and other available data sets and the result is a synthetic population with average statistics similar to the actual one by design, to the point that one is not able to tell if an individual belongs to the real or to the synthetic population, with the advantage of being relieved from most privacy concerns.
In this context, we have generated the synthetic population of France, based on Census data from INSEE (French Institute of Statistics and Economic Studies) and other data sets available.
The main data sets involved: information at individual level, such as age, sex, level of education, household composition, etc.; information at household level, such as sociodemographic characteristics as well as information about the dwelling and its location, characteristics, category, type of construction, comfort, surface area, number of rooms etc.; information about mobility to the workplace, including their main socio-demographic characteristics, as well as those of the household to which they belong; information about the mobility to education facilities.
Additional datasets included the map of the census tracks used by the INSEE, and data from the cadastre about properties cross linked with geographic data form the French Geographic Institute (IGN) and OpenStreetMap to create as detailed a map with the distribution of dwellings by type. Data about the location of educational establishments was extracted from the Ministry of Education, while the location of economic activities was obtained by cross referencing the data from INSEE which covers 64 different economic activities with the buildings for the OpenStreetMap database.
By linking the datasets above it was possible in the first instance to create families and households and then to attribute them to individual buildings. This combinatorial optimization is known as the Variable Size Multiple Knapsack Problem. This problem can be tackled in different ways, no solution is perfect but there is always a trade-off between precision and computational intensity. Aiming at a better precision is only possible when the input data adds up useful information. Sometimes, the least computationally intensive solutions offer reasonable results as well. In our case, having any additional attribute to houses, e.g. year when built, would make people positioning much more precise. Another source of uncertainty is that, in the absence of better information, we assumed that larger families would inhabit larger housing surfaces, which is obviously not always the case.
Notwithstanding these limitations, we modelled the synthetic population of 63 million people, in 35 million

households allocated in 10 million houses in France including their travel to work and study places behaviour.
The computations were performed in batch processing on the JRC Big Data Analytics Platform (BDAP), that uses HTCondor as a job scheduler and Docker Universe set up.
Around 35k jobs were performed, one for each French commune, each job taking 1 CPU. At our disposal were 20 servers of 40 CPUs each and 1TB RAM, and relatively unlimited storage space. The machine set was shared with other users.
The scripts were in Bash and Python, including libraries such as Numpy, Pandas, geoPandas and Shapely.
One of the challenges was to deal with very large CSV files in input (e.g. one of 12GB). Opening these files (in Pandas) required that the memory demand in the Condor submit file had to be so large (˜200GB) that machines were seldom allocated to our jobs.
The idea was to subset from the large files only the records that belong to the job that is performing, so to make a query for a certain value (zip code processed by the job) along a certain column (zip code column), and subset only those lines that correspond to that query and save the result in a new CSV.
A benchmark of several libraries used for subsetting was performed and eventually the winner was AWK, offering the best speed and inferior memory requirement.

Speaker release

Desired slot length

Primary authors: DI LEO, Margherita; HRADEC, Jiri (Joint Research Centre)

Presenter: DI LEO, Margherita

Session Classification: Workshop session

Track Classification: HTCondor user presentations