KPI for geolocalised services

DAF Analysis Group

May 2018

1 Introduction

This short paper presents a simple way to define KPIs for the geolocalised services in a city at the neighbourhood level, in order to express how well the services availability meets the local demand.

Let us suppose that for a certain city we have geolocalised data for different service types, such as schools, libraries, pharmacies, public transport facilities and parks.
Each of these categories has several units, that might differ either in the population age ranges they target, as in the case of a primary school versus a secondary one, or in their capacity.
We assume that these services have to be visited by citizens in order to be used.

Suppose we also know with good approximation (roughly one block) where the population lives and the corresponding distribution of the ages. This kind of population data can be found at the "census sections" level ("Sezioni di Censimento"), the finest-grained unit defined by the national statistical system.

In the following we present a simple computational model that allows to estimate the match between demand and offer of public services in a city.[1] The aim of its development is to provide domain experts with an example of an open source model that can be easily expanded and maintained.

What we look forward to is that domain experts (e.g. civil servants in the local municipalities) can partner with data scientists to build tailored models. These maintainable models avoid manual bespoke processing for each run and provide reusable outputs for policy-making and transparency goals.
In this way, local institutions can engage in higher added-value activities and continuously monitor the impact of their policies.

[1] This is in general a quite complex problem. If we consider schools, for example, an interesting working paper by Dinerstein and Smith can be found at *http://economics.mit.edu/files/11164/*.

2 Model

Let us consider a generic service, for example schools, located at $(P_i)_{i=1}^{I}$, that serve a set of census districts ("Sezioni di Censimento") represented by their centroids at location $(Q_j)_{j=1}^{J}$ and having population $(m_{wj})_{j=1}^{J}{}_{w=1}^{W}$ where I is the number of schools, J is the number of sections and W is the number of age groups in which the population is stratified.

2.1 Location-based supply modelling

In what follows we will assume that the service we are considering is represented by schools, but the same reasoning can be applied to other types of services (for example libraries...). We describe the supply of school i for age group w at location Q with a *radial basis Gaussian kernel* centered at P_i:

$$k_{wi}(Q)$$

and a single lengthscale parameter (the standard deviation of the Gaussian distribution). If a school (or a generic service unit) has a capacity, we scale the lengthscale with the capacity, otherwise we set the lengthscale to a conventional value (e.g. 0.5 km).

The supply of school i for age group w at section j is then:

$$k_{wij} = k_{wi}(Q_j).$$

2.2 Demand modelling

We assume that the demand of service i for age group w at district j equals the number of residents in district j that belong to age group w, i.e. m_{wj}.

2.3 A utility function based on supply and demand

From 2.1 we know the supply of each unit i (e.g., a school) at each demand location Q_j and from 2.2 we know how many people live at Q_j.

We define a utility function by re-scaling the estimated supply k_{wij} with a_i, what we call *the total attendance* at service unit i:

$$\widetilde{k}_{wij} \propto \frac{k_{wij}}{a_i}$$

under the assumption that lower attendance is always better. The total attendance is defined below.

We define a_i, the total attendance for service unit i, as:

$$a_i = \sum_{w=1}^{W} a_{wi},$$

the sum over the different age groups w of a_{wi} where a_{wi} is *the attendance* for service unit i and age group w and it is equal to the weighted sum over all districts j of the demand m_{wj} at district j:

$$a_{wi} = \sum_{j=1}^{J} \frac{k_{wij}}{\sum_l k_{wlj}} m_{wj}.$$

The weight $\frac{k_{wij}}{\sum_l k_{wlj}}$ is the relative service level of unit i in district j, when compared to the alternatives that are available in district j. For example, if a specific age group at Q_j has 3 alternatives with kernel values of $(0.6, 0.6, 0.8)$, then we will assign $\frac{0.6}{0.6+0.6+0.8} = 30\%$ of m_{wj} to unit 1, 30% to unit 2 and 40% to unit 3. Under-served areas have high attendance while well-served areas have low attendance, under the assumption that at each demand location Q_j, residents in each age group will use the service units in the city *proportionally to* (or at least with an increasing law of) the computed service levels.[2]

As a final note, in the definition of the utility above, we assumed that lower attendance is always better. For some kind of services, like social places or sporting activities, this assumption might not be valid, and a different model for the utility function might be used, or the same model with additional parameters.

2.3.1 Normalization of the utility function

To compare utilities, we define a constant of proportionality a_i^{ref}:

$$\widetilde{k}_{wij} = \left(\frac{a_i^{ref}}{a_i} \right) k_{wij}$$

which, depending on the specific service type we set equal to:

1. The specific capacity???, if we have data about service unit capacities (e.g., school capacity or park size);

2. The *observed mean attendance*, if we don't have such data. When applying this model to different cities, we use the same reference level for comparable services not to allow for city-specific effects.

As the attendance should rescale the service supply levels within some bounds (e.g., if a pharmacy is used very little, its service supply should not raise indefinitely) we multiply the utility function by a non-decreasing function f of the

[2]This excludes the possibility for agents to consider the attendance dynamics in their evaluation. In case we were to deal with interactions, a suitable framing would be a congestion game setting. In a congestion game, players needs to decide how to consume a limited set of resources without the ability to coordinate; they typically get less utility if the resources they decide to use are also chosen by many other players. A common example is choosing the fastest route in a road network. The congestion game setup can definitely be solved (at least in an approximate way, see [1]) but requires more detailed inputs about preferences and types in the population.

relative attendance $\frac{\bar{a}_S}{a_i}$.
A possible simple choice for f can be a clipping rule with a threshold factor L:

$$f_L(x) = \begin{cases} 1/L & \text{if } x < 1/L \\ x & \text{if } x \in [1/L, L] \\ L & \text{if } x > L. \end{cases} .$$

We choose $L = 1.4$ to allow for a maximum attendance correction of 40% on the original service supply levels.
The final kernel functions corrected for attendance are then:

$$\widetilde{k}_{wij} = f_{1.4}\left(\frac{a_i^{ref}}{a_i}\right) k_{wij}.$$

2.3.2 Aggregation at the census-section-level

To get an index of the service quality for age group w at Q_j, we can use a vector norm to aggregate $\widetilde{k}_{wij}$ on index i.
A default choice for the aggregation norm is the l2 (Euclidean) norm, that gives a less-than-linear premium for having several service units available:

$$v_{wj} = ||\widetilde{k}_{wj}||.$$

2.3.3 Aggregation at the neighbourhood-level (KPI)

At this point, we have an estimate v_{wj} of how the service supply matches the demand of each age group w in each census section j, i.e., at the census section level. To get an estimate $v_{w\hat{J}}$ of the service supply at the neighbourhood level, we aggregate local estimates using a *social welfare function*. Here index $\hat{J}$ represents a neighbourhood and aggregates the index of all sections j belonging to it.

A naive choice for the social welfare function is given by the utilitarian welfare function, i.e., the weighted average of the utilities:

$$v_{w\hat{J}} = \frac{1}{\sum_{j\in\hat{J}} m_{wj}} \sum_{j\in\hat{J}} m_{wj} v_{wj}.$$

Alternatives to the utilitarian welfare function that are more representative of neighbourhood inequality are measures based on the Gini coefficient, on entropy, or on the Rawlsian function[3] which defines the social welfare function at the neighbourhood level using the minimum utility level in the neighbourhood

$$v_{w\hat{J}}^{\text{Rawl}} = \min_{j\in\hat{J}} v_{wj}$$

thus considering only the worst value in the neighbourhood.

In principle we could use different social welfare functions and compare results.

[3] Amartya Sen proposed $\bar{Y}(1 - \text{Gini})$; other similar measures involve Thiel-L index for the distribution.

3 Bibliography

[1] Vazirani, Vijay V.; Nisan, Noam; Roughgarden, Tim; Tardos, Éva (2007). *Algorithmic Game Theory*. Cambridge, UK: Cambridge University Press.