

Background

In the Swedish CPI, the sample of outlets for manual price collection is a probability sample of approximately 800 outlets, stratified by industry. Up until 2015, geographical location was not taken into account.

Manual price collection in stores is expensive and travel expenses account for about 25% of the total cost. In an attempt to reduce travel, the 2015 sample of stores was concentrated to a probability sample of geographical areas. Stores are then sampled within those areas. This early effort did have an effect on overall cost, since total travel time was reduced.

We believe that it is possible to make the geographical sample even more cost efficient by utilizing information on price collector's home addresses and ease of travel.

Two-Phase Sampling of Outlets

Phase 1: Geographical Sample

- Systematic π ps sample of *postal codes* with inclusion probabilities proportional to total trade in each postal code area.
- Geographical Cutoff
 - Must cover at least 90% of total sales in Sweden.
 - Some hard to reach areas are hand picked for cutoff.
 - Municipalities are ranked according to total sales and the smallest on that list are picked to complete cutoff.
 - Total sales in a cutoff area is compensated for on county level, by increasing the inclusion probabilities for remaining outlets in the the same county.

Phase 2: Sample of Outlets

- Sequential Poisson sample of outlets with inclusion probabilities proportional to total size multiplied with the inverted probability from phase 1.
- Size is a function of total turnover and number of employees per outlet, adjusted to represent private consumption only.
- Within industry, the sample of outlets is balanced to prevent a disproportionate concentration of stores to any geographical area.

Towards a More Efficient Design

By using information on where price collectors live, it should be possible to create an even more cost efficient sampling design. Since travel cost is high, the main objective must be to sample outlets close to price collectors.

- Which geographical division should be used for phase 1?
- Can we use road maps to optimize traveling time for price collectors?
- Do we need to review the distribution of price collectors?
- We see a clustering effect in variances
 - Do stores in close proximity have similar price changes?
 - Is it a price collector effect?

Clustering Effects - a Closer Look

In the 2017 sample, there seems to be a correlation in price movements between outlets in close proximity. Comparing variance structures, we see that within an industry, price movements in outlets within the same area are more alike than those in outlets in different areas. This effect is stronger within smaller areas.

- *Postal Codes*
 - Geographical area used for sampling today
 - Not an administrative area - used by the postal service
 - Does not sum to other administrative divisions, e.g. municipalities
 - 10 000 postal codes in Sweden with retail trade
- *DeSO*
 - A division created by Statistics Sweden to use for demographic statistics, DeSO is short for demographical statistical area in Swedish
 - Sums to Sweden's administrative divisions
 - 5 985 DeSo in Sweden
- *Municipalities*
 - 290 municipalities in Swedens

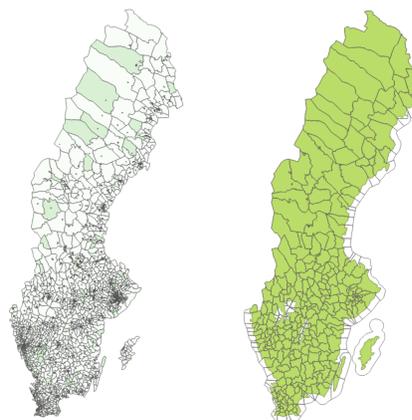


Figure: To the left we see the geographical division DeSO, to the right municipalities.

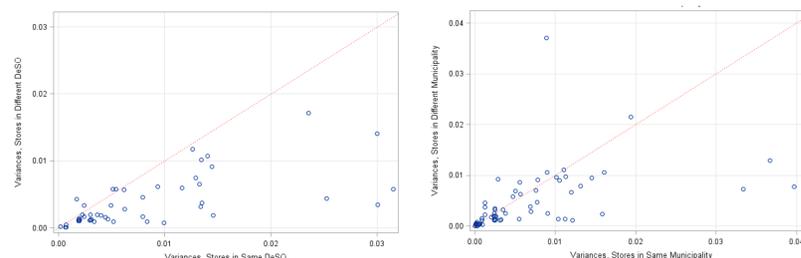


Figure: Variance structures within and between DeSo to the left, and within and between municipalities to the right.

A Possible Solution?

Balanced sampling resembles quota sampling in that control variables are used to guide the sample selection. The objective then is to select a sample such that the sample mean of each control variable equals (or closely approximates) its known population counterpart.

By balancing our sample we can make sure not to pick two outlets from the same industry in the same geographical area.

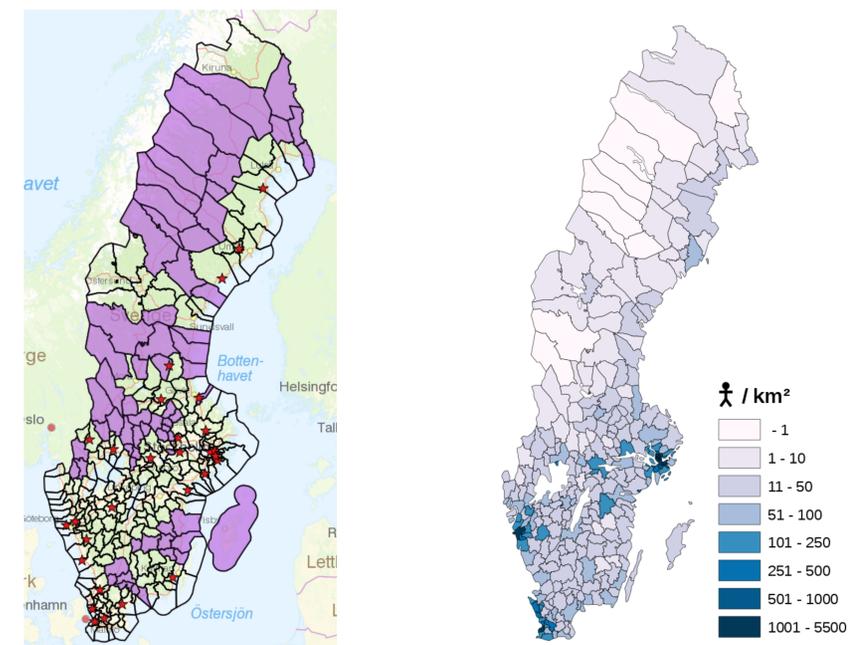
Historically, a sample balancing scheme has been applied to the Swedish sample of outlets (see Phase 2: Sample of Outlets). With a more thorough understanding of the variance structures in the population and with clear set constraints on travel, an improved balancing scheme should be possible.

Geographical Cutoff

The cutoff applied to the sample in 2017 was derived solely from information about total trade per area. In the left map below, cutoff areas are marked with purple and every star represents a price collector. It is obvious that we need to consider where price collectors live in order to create a cost efficient design.

Since 2017, the data collection unit has undergone several changes and the number of price collectors has declined. Starting with the sample for 2019, the cutoff will be derived using price collector information as well as information of total trade.

Below to the right is a map showing the population structure of Sweden. Areas with low population density often have less trade than areas with high population density. In addition, it is in general more difficult to recruit price collectors in rural areas, especially in the northern part of the country.



In designing the cutoff, it could be preferable to consider total travel *time* rather than travel *distance*. This is of course more involved, but may prove to be well worth the effort.

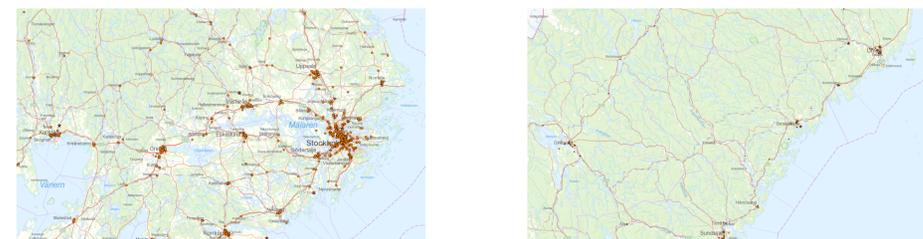


Figure: Distance is not everything. In rural areas in particular, the closest distance does not always mean the easiest travel. Örnsköldsvik would be easier to reach from Sundsvall than it is from Örnsköldsvik, especially during snowy winters.