Mountain Doodles

Geocoding errors in aggregate data

Jens von Bergmann MountainMath

Geocoding errors can become a problem, especially in temporal data where geocoding processes have changed over time. We propose a method to detect and ameliorate geocoding errors in aggregate data.
This method is useful and effective in detecting and dealing with geocoding problems where data gets misattributed to neighbouring regions. Addressing this is important as these errors tend to happen precisely in regions that have experienced significant change and thus correlate with changes in density, which has a high chance of biasing analysis. The process of joining regions across boundaries that were subject to geocoding problems solves this problem in a similar way that Tongfen solves the problem of changing (census) geographies across time, and most importantly it does so in a way that avoids biasing the analysis. This solution comes at the price of losing control over the exact geographic regions used by resulting in a slightly coarser analysis.
The method relies on a set of parameters that allow to fine-tune the process and trade risks of missing cases where data was erroneously geocoded to a neighbouring region against being overly cautious and removing too much geographic grannularity in the data.

2024-07-26

Determining the exact location of data points comes with measurement errors, in particular when geographic locations is derived from geocoding. Data that gets aggregated by geographic areas relies on geocoding to determine which area each data point belongs to.

Geocoding is the process of converting addresses into geographic coordinates, which can then be used to determine which area each data point belongs to. However, geocoding is not perfect, and the resulting location can have errors. When aggregating data to polygons those errors can lead to data points being assigned to the wrong polygon.

If all data was geocoded with the same process, as for example using single year census data, this does not matter much, it might just shift the relative composition of variables at small geographies but data associated with housing and the people therein is kept intact. But when looking at multi-year census data, with data being geocoded with separate processes in each census year, this can lead to problems.

As an example, consider the 1971 through 2011 semi-custom tabulation on 2016 dissemination area geographies (Statistics Canada 2019) that we have worked with extensively in the past. StatCan geocodes building by *block face*, which is based on their road network file with attached block level civic numbers. In some areas, especially when roads did not exist in earlier census years or were substantially changed since then, the geocoding was done inconsistently and will lead to counts being incorrectly attributed to neighbouring areas in some cases and census years. One solution to this is to join neighbouring regions across which data has been misattributed, resulting in a slightly coarser geography but with consistent data over time.

Tongfen

Before we dive into details and examples, we want to situate this problem within a broader set of problems of making data comparable through time. Our Tongfen package addresses a large set of use cases where temporal data comes on different yet congruent geographies by creating a least *common denominator geography* and re-aggregating data accordingly. (von Bergmann 2020b) This stands in contrast with other approach that estimate data based on areal or dasymetric interpolation that has been shown to lead to significant errors. (Logan, Stults, and Xu 2016) This becomes problematic because the errors typically correlate with changes in density, which in turn often correlates with variables of interest, which can introduce systematic bias in analysis. (von Bergmann 2020c, 2020a)

Our current problem faces similar issues in that inconsistent geocoding over time is often associated with significant changes to the built environment, and thus the very regions we are most interested in and that are likely be responsible for a significant portion of effect sizes in analysis.

We think of the correction of geocoding errors by joining adjacent regions as an extension of the Tongfen functionality and will, after some more testing and refinements, incorporate this into the package.

Census timelines

The semi-custom tabulation of census profile data for Toronto and Vancouver to 2016 dissemination areas will serve as our main example. In Figure 1 we replicated the work from the previous post

to take the 1971 through 2011 cross tabulation on 2016 dissemination areas, augment it with the 2016 data, and use TongFen to extend it to 2021. (von Bergmann 2024)

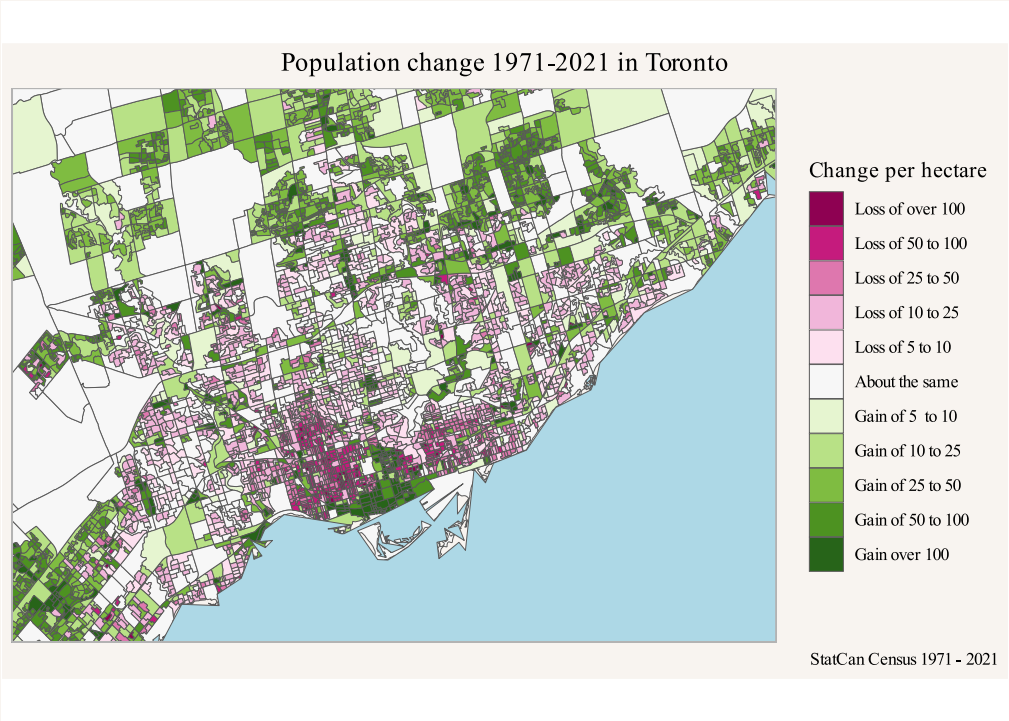

Figure 1

The interactive map makes it possible to browse the data in more detail. And it makes it possible to identify quirks likely due to geocoding problems. But what do these geocoding problems look like? Time for a concrete example.

Crescent Town

A particular example, which also provided much of the motivation for this work, was flagged to me by Tai Huynh from The Local, who has been using this data to look in more detail at neighbourhood change in Toronto. Here the population jumped in erratic ways in the early census years as visualized in Figure 2.

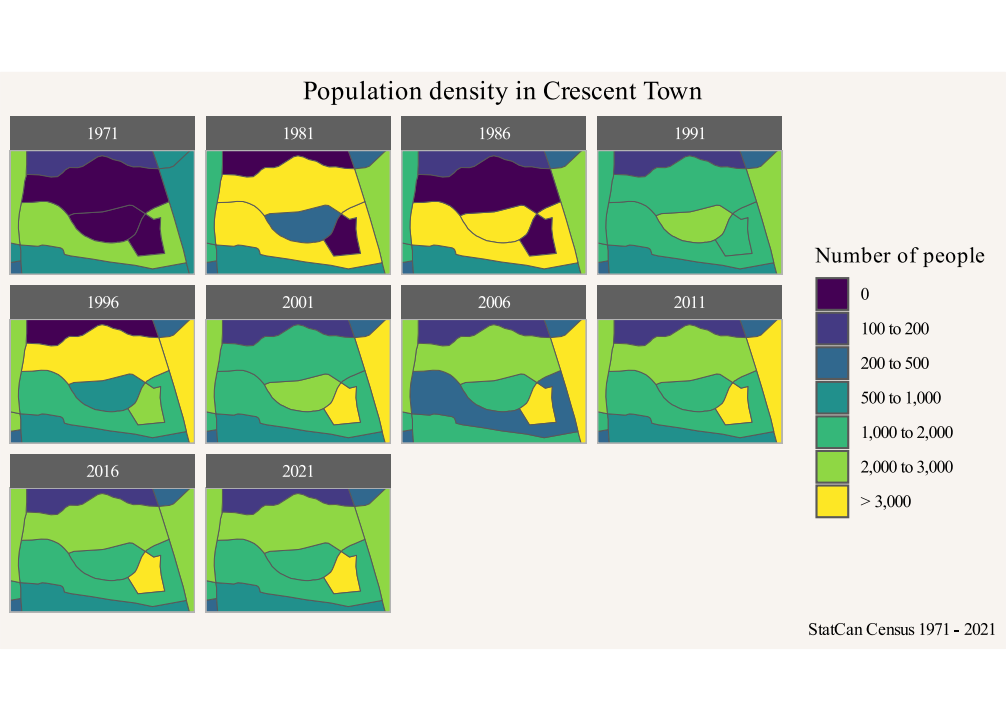

Figure 2

We see the central moon-shaped area starting out with zero population in 1971, growing to very high population in 1986, but then dropping back to low levels in 1996. At the same time, some of the surrounding areas are showing opposite patterns.

In Figure 3 we highlight the population timelines of the moon-shaped central area and the surrounding ones in a line graph.

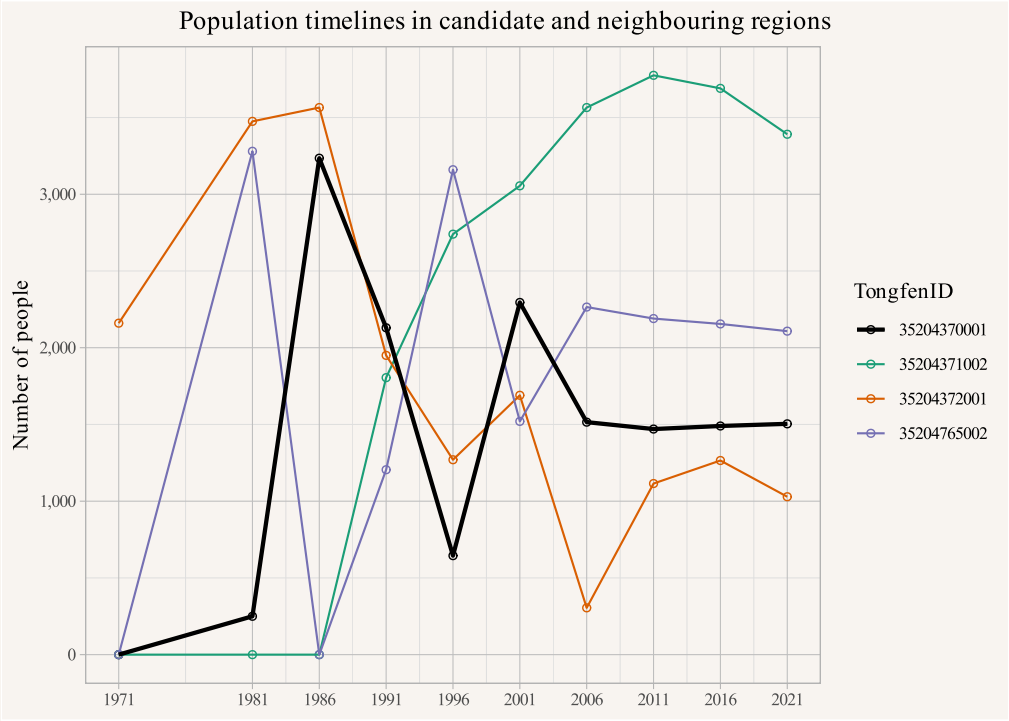

Figure 3

Figure 4 shows the regions shown here on a map for added context.

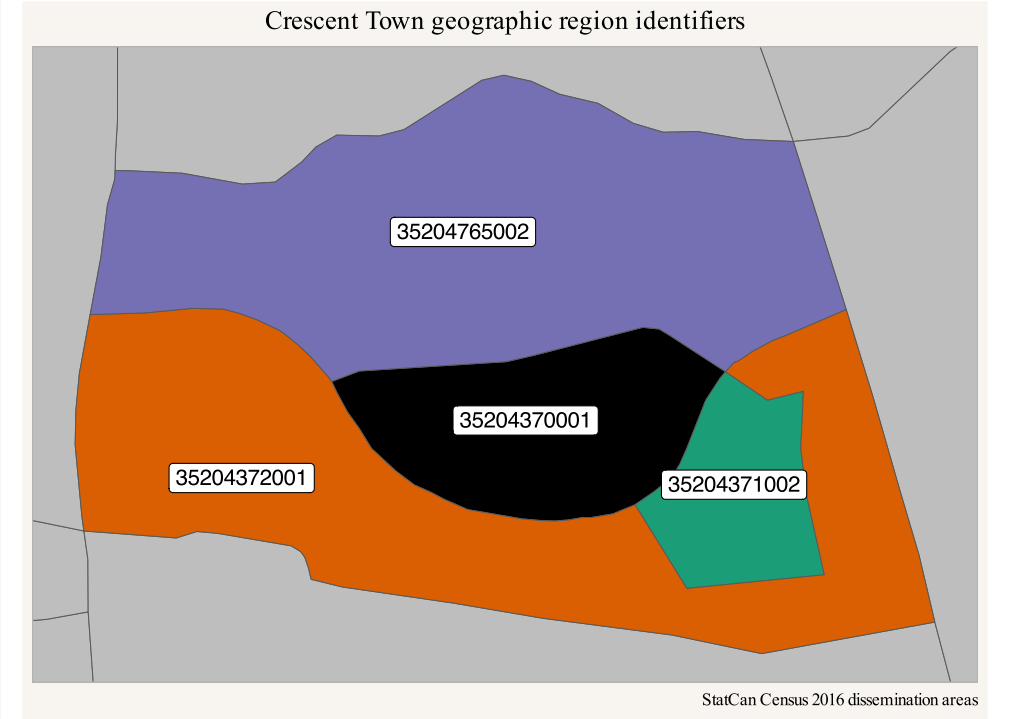

Figure 4

The erratic behaviour of the timelines in the individual regions is quite striking, especially regions with zero population at the start point to development in previously undeveloped (for housing) areas that might come along with reconfiguration of streets that trigger geocoding inconsistencies, and population dropping from high values to zero is a red flag. But if we squint long enough at the data we notice that among those regions there are two pairs with complementary and partially cancelling patterns. Figure 5 shows what happens when we combine join the region, and the corresponding populations.

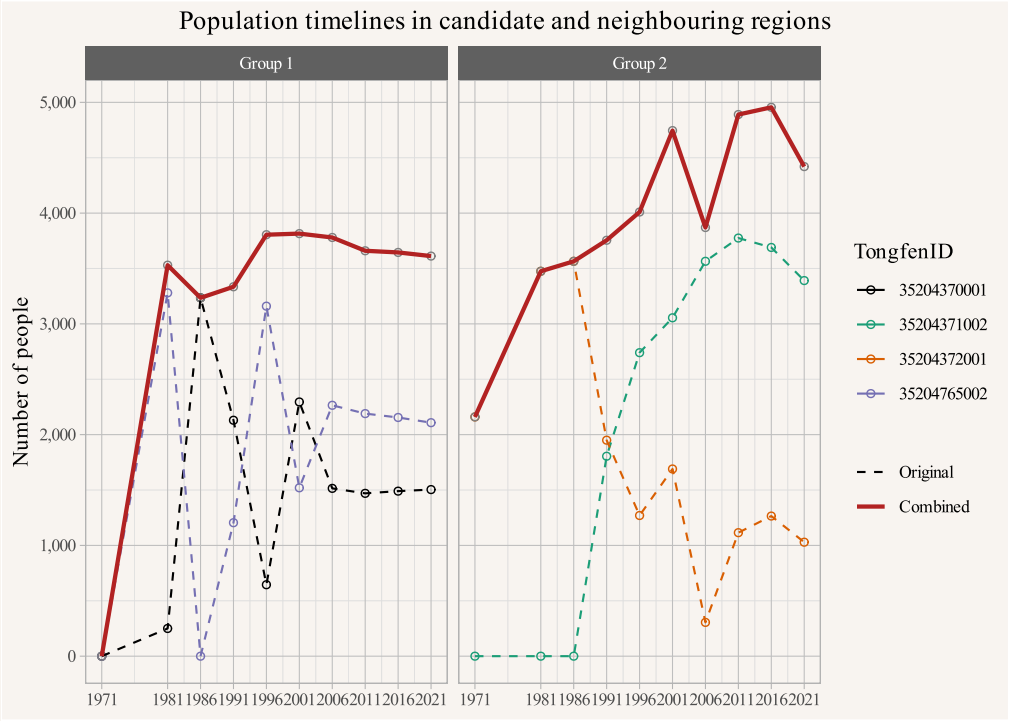

Figure 5

The lines in red showing the populations of the combined regions are much more stable and don't exhibit the drastic population changes that the individual regions do. This gives us fairly high confidence that the dramatic population changes in the individual regions were caused by geocoding errors, and that combining the regions fixes this (at the expense of moving to a slightly coarser geography). Figure 6 shows the resulting timeline on the combined geographic regions.

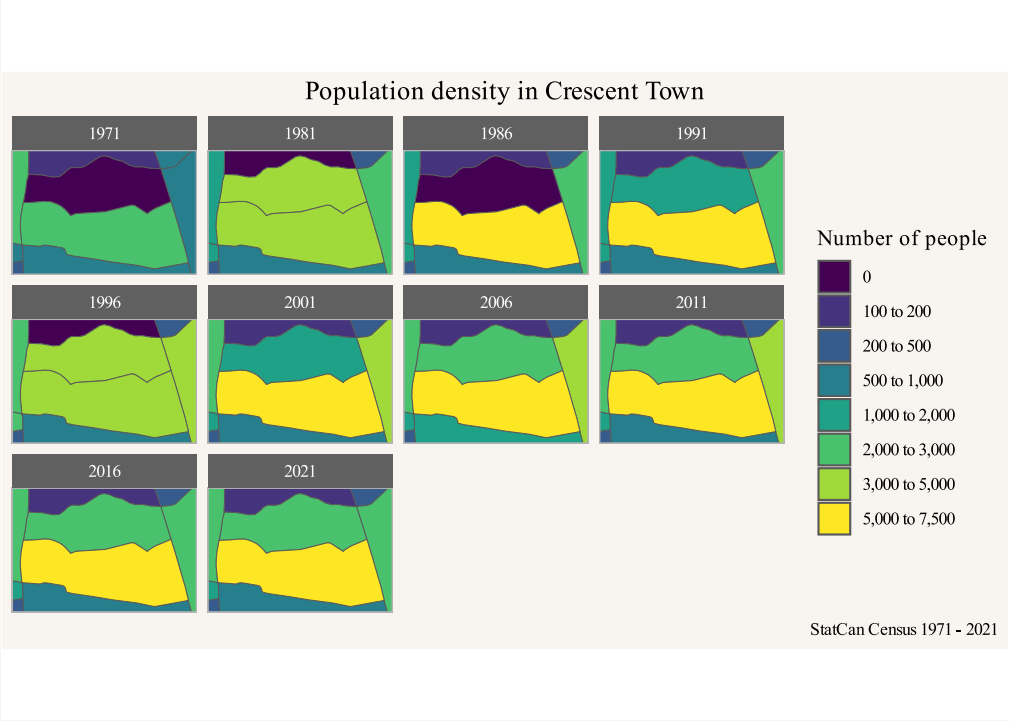

Figure 6

Hopefully this example has made the problem clear, and also shown, at least in principle, how to fix this problem. But this becomes a daunting task when we have to do this for all 7,518 regions in Metro Toronto. We need a way to automate this process.

Automating the process

Automation comes with risks of missing regions that should be combined, and on the flip side it risks of being too aggressive and joining regions with suspicious population changes that aren't due to geocoding but to real processes on the ground. And these risks vary by region and by type of data considered. We need a process that is flexible enough to be adapted to a range of situations, and that can be calibrated based on use cases and tolerance for type 1 vs type 2 errors.

Identifying potential geocoding problems

As a first step we have to draw up a list of candidate regions that might have geocoding problems. Intuitively we are looking regions with "surprising" jumps in population. But what makes a population change "surprising"? We have looked a similar problem before (von Bergmann 2017), to determine surprise we have to first build an expectation of what the population change might look like and then look for significant deviations from this expectation.

Do build and expectation of population change we have to understand the data generation process. Generally, we expect population change to be gradual, with possibly large increases when new residential development occurs. Strong decreases in population are less common, but they could occur e.g. during a large scale fire or natural disaster. Based on this we formulate our expectation fairly broadly as "significant population decreases over relatively short periods of time are surprising". This will constitute our first round of candidate areas for geocoding problems.

What constitutes "significant" or "surprising" change is context dependent, Figure 7 gives an overview of the distribution of the negative absolute and relative population changes in the data.

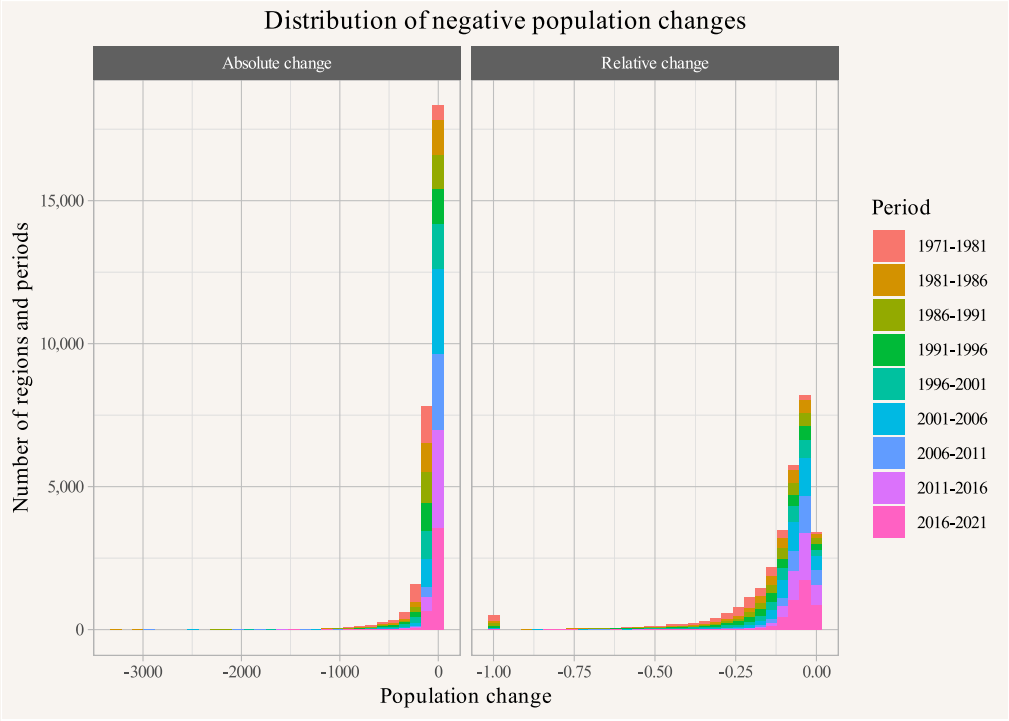

Figure 7

Based on this we somewhat heuristically and conservatively take a population drop of at least 100 people and at least 20% between two census periods as starting to be a "surprising" change and formalize this by modelling the amount of "surprise" with an inverse exponential distribution with a "surprise" value of 0 at 0 population change and a "surprise" value of 0.5 at 0.25 and 100 relative and absolute population change respectively. We then estimate the total surprise by multiplying the relative and absolute surprise values, with our candidate regions being those with a total surprise of at least 0.25, so just capturing a case where a region's population decreased by 0.25% and 100 people. Figure 8 shows the distribution of total surprise.

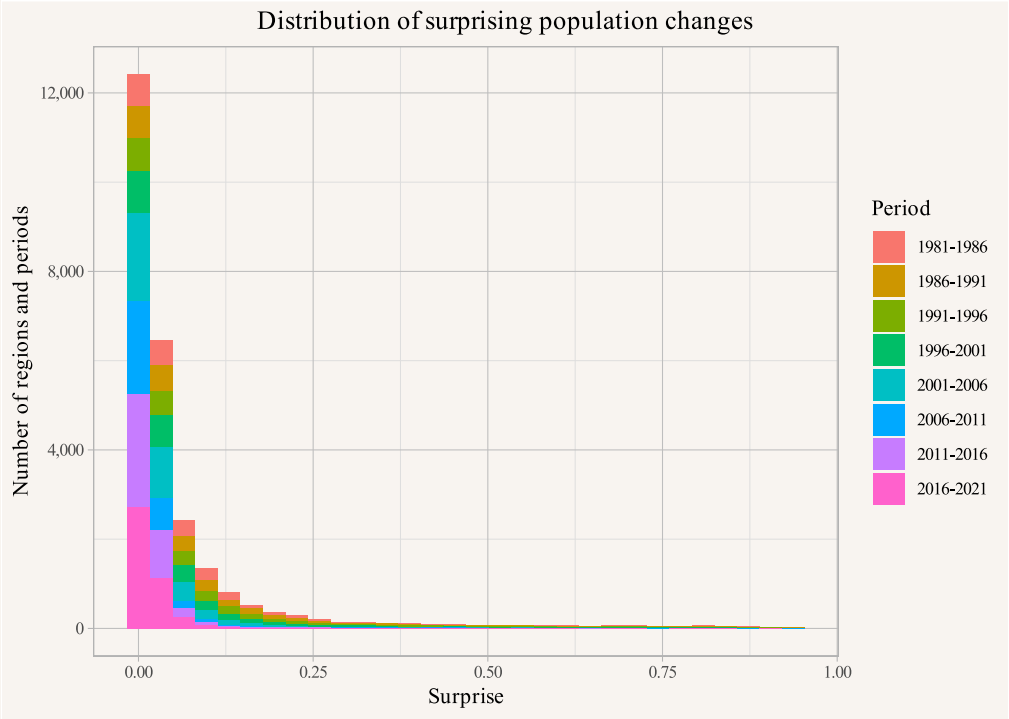

Figure 8

To get a better view of the regions we are most interested in we zoom in a bit and only show regions with a total surprise of at least 0.15 in Figure 9.

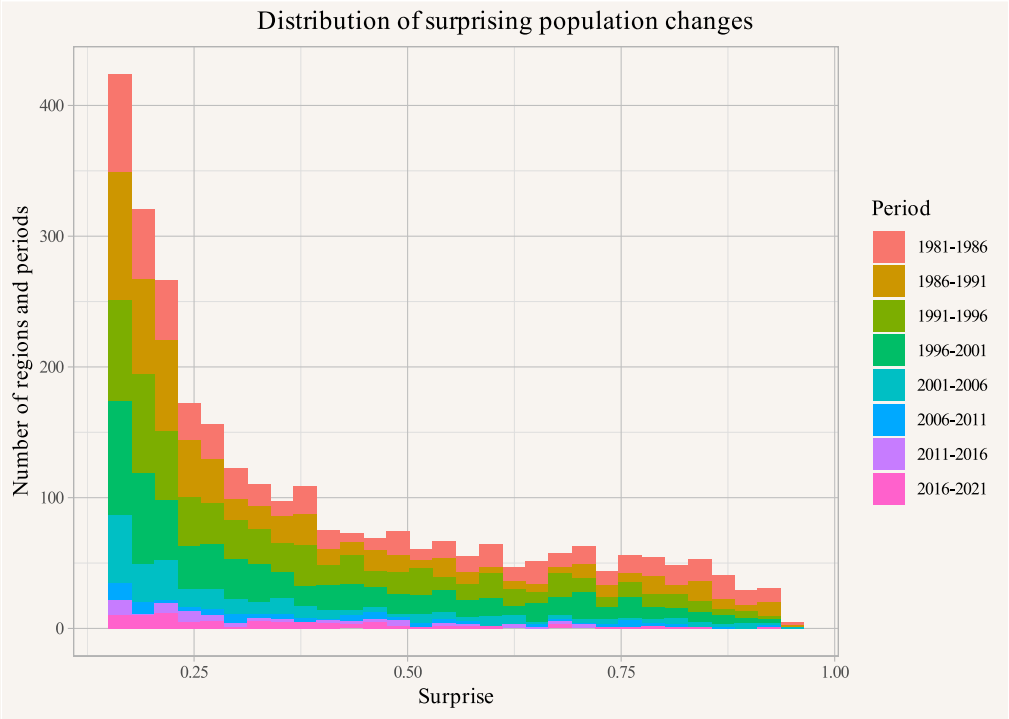

Figure 9

We note that the candidates lean heavy toward the earlier periods, which is expected when looking for geocoding errors, but also consistent with this being a time period of slower growth, and possibly stronger declines in some areas that reflects real conditions on the ground. But as mentioned above, not all these surprising population changes are due to geocoding errors. And geocoding errors can also lead to strong increases in population. To account for this we take the first round candidates and consider the neighbouring regions for each of these to look for complementary patterns with population boosts in years where we see surprising population drops.

To implement this we make this more precise. Ideally we would look at the joint distribution of pairs (or higher order combinations) of regions and estimate the surprise and reduction in surprise at that level, but for exploratory as well as practical considerations regarding the computational intensity of such an approach we settle for a more pedestrian approach. At this point this is still somewhat heuristic, we plan on refining this and settling on a principled and parsimonious approach. At this point, we proceed as follows. We rank all regions by *total surprise* by taking the L^p norm across years, with p chosen to be 4 in this case, but retaining the ability to tune this further in order to trade off between only considering the maximum surprise across years ($p = \infty$) and simply taking the sum of all the individual surprises ($p = 1$) across years.

Guided by our example case we will look for neighbouring regions that, when combined with the candidate region, result in a significant drop in our *surprise* metric when joining the regions. To estimate the reduction in surprise we remain mindful that surprise decreases when we simply add a

region with flat population as that increases the denominators in our relative surprise measure. To avoid this effect and measure the reduction in surprise in a comparable way we keep the original denominators but the take the combined change in population as new numerators when evaluating the change in surprise. When then combine the regions if the relative or absolute reduction in surprise passes a specified threshold that we deemed significant enough. We repeat this process until we don't see any significant reductions in surprise anymore.

The automated process performed 976 region joins, reducing the total number of regions from 7,518 to 6,542 to arrive at a slightly coarser geography.

Checking back in with our motivating example of Crescent Town we see in Figure 10 that the automatic process resulted in the same grouping that we had identified manually.

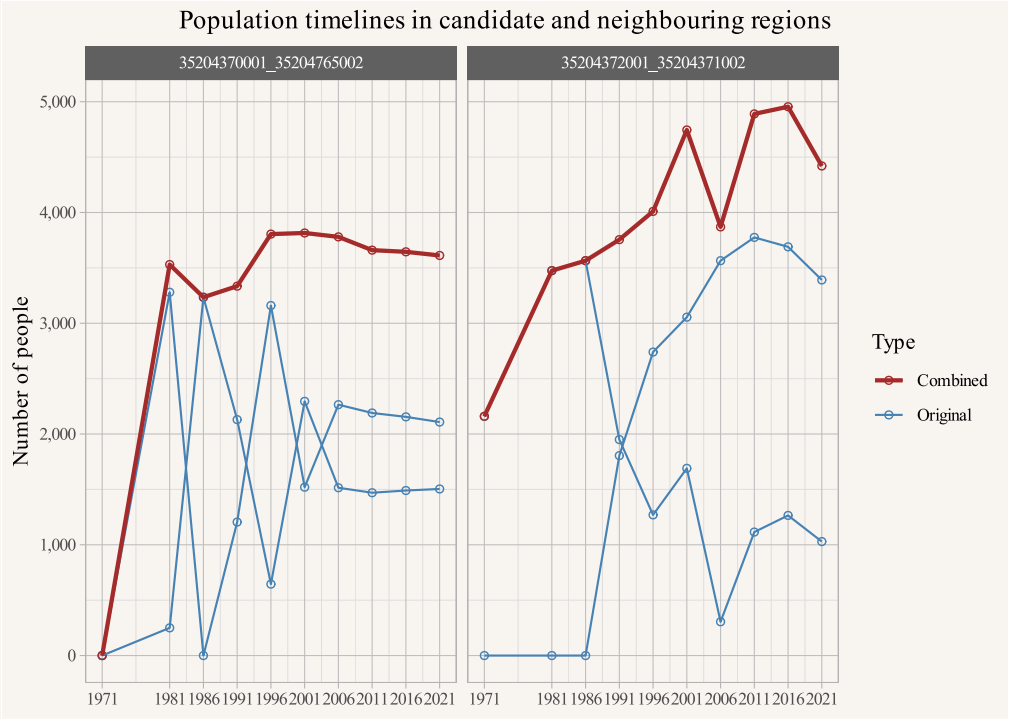

Figure 10

Vancouver version

We can perform the same operations on the Vancouver data and identify possibly geocoding problems there, the resulting cumulative population change is shown in Figure 11.

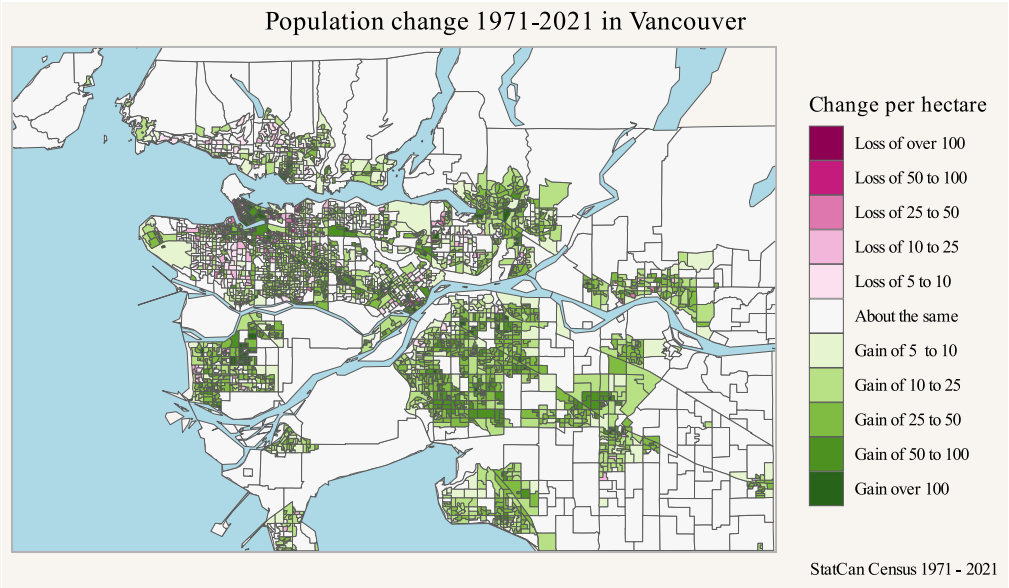

Figure 11

Based on this we again look for surprising population changes, the distribution shown in Figure 12 looks quite similar to the Toronto version.

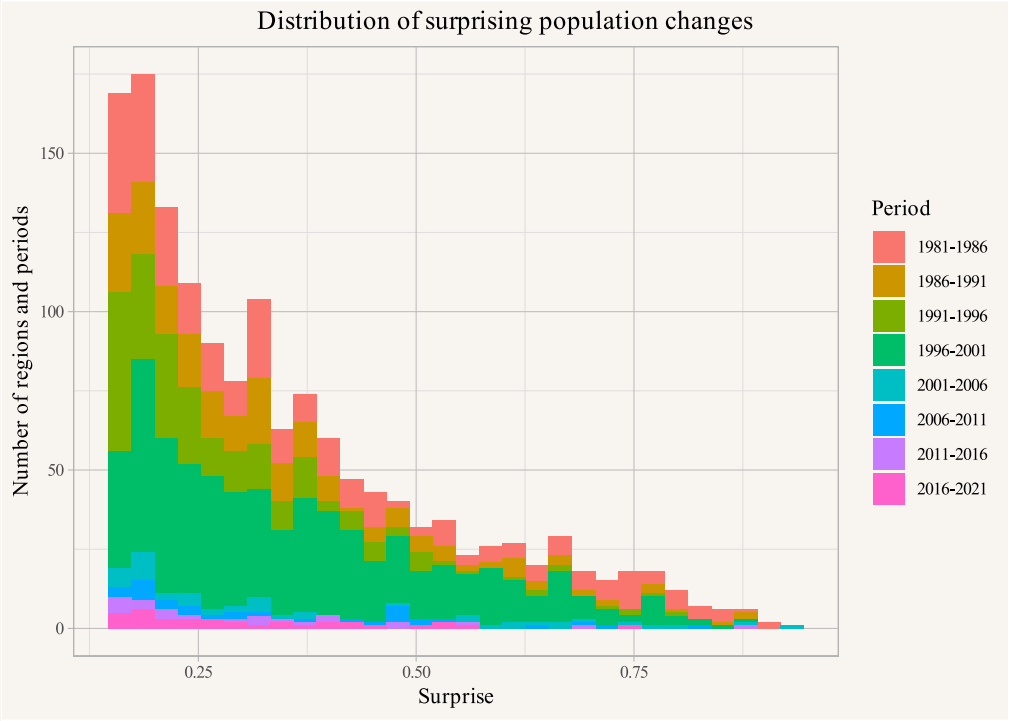

Figure 12

Repeating the procedure with the same parameters as in Toronto results in 416 region matches and correspondingly slightly coarser geographies.

Once this is automated it's simple to apply this to other regions or other types of data. But it's generally a good idea to take a more detailed look at how it performs on some examples.

Little Mountain

One danger of employing this procedure is that we might lose information of localized population declines that are real. Strong real population declines are rare, population change tends to be slow and dramatic declines rarely manifest themselves over a single 5 year inter-census period. But there are some examples where we can see *surprising* population declines that correspond to real process, Little Mountain in Vancouver provides a prominent example, with plenty of good news coverage with background information. This was a social housing development that got sold off by the province to private developers and most tenants were evicted in 2009 and buildings were subsequently torn down, with planned replacement housing on the site mostly stalled since then. This lead to a *surprising* drop in population in those areas between the 2006 and 2011 censuses as can be seen when looking at census timelines of the two corresponding dissemination areas. Figure 13 shows the population change 2006 to 2011 in the two regions covering the Little Mountain development, as well as the surrounding regions.

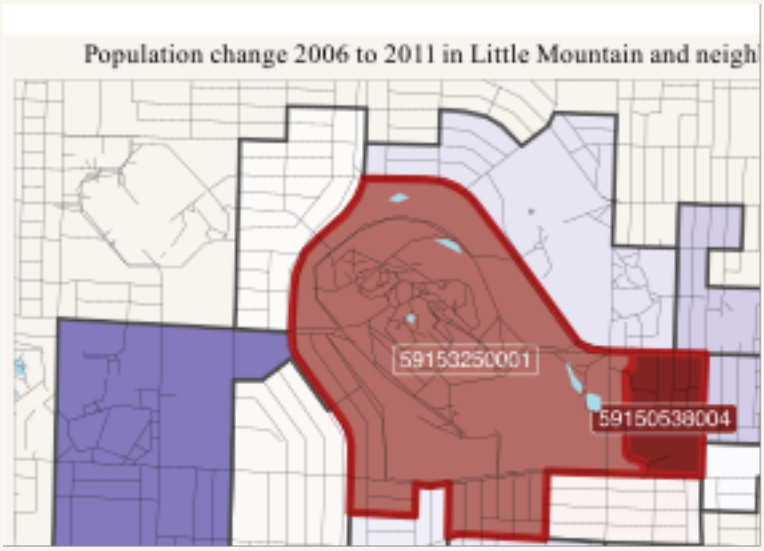

Figure 13

The two areas that overlap the Little Mountain site also include neighbouring regions, but the drop in population is noticeable for both.

Figure 14 shows the full timelines of the regions, we note that neighbouring regions don't exhibit complementary patterns that suggest that this drop was due to geocoding errors.

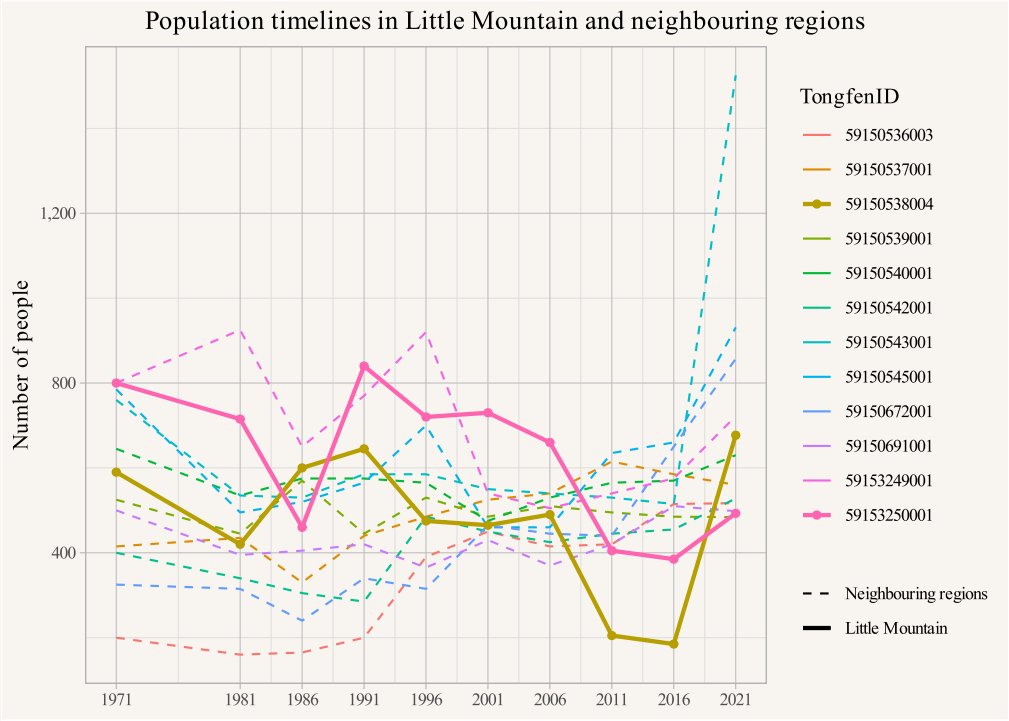

Figure 14

While the algorithm does flag the regions as *possible candidates* for geocoding errors, it does not find enough evidence that this is due to patterns consistent with geocoding problems because of the lack of complementary patterns in neighbouring regions. Considering examples like this can help calibrate the sensitivity of the algorithm.

Conclusion

The current implementation is useful and effective in detecting and dealing with geocoding problems where data gets misattributed to neighbouring regions. Dealing with this is important as these errors tend to happen precisely in regions that have experienced significant change and thus correlate with changes in density, which has a high chance of biasing analysis. The process of joining regions across boundaries that were subject to geocoding problems solves this problem in a similar way that Tongfen solves the problem of changing census geographies across time, and most importantly it does so in a way that avoids biasing the analysis.

The downside is the loss of control over the exact geographic regions used to aggregate the data and a resulting slightly coarser geography. In applications where it is important to retain the given geographies one has to find other solutions to solve this problem.

In the near future we plan to integrate this functionality in the Tongfen package to make it easier to use, and also to tie into existing functionality to re-aggregate data on joint geographies.

As usual, the code for this post is available on GitHub for anyone to reproduce or adapt for their own purposes.

Logan, John R., Brian J. Stults, and Zengwang Xu. 2016. "Validating Population Estimates for Harmonized Census Tract Data, 2000–2010." *Annals of the American Association of Geographers* 106 (5): 1013–29. https://doi.org/10.1080/24694452.2016.1187060.

Statistics Canada. 2019. "1971, 1981-2011 Census Profiles and 2011 National Household Survey Profile Aggregated to 2016 Dissemination Area Level Geographies [Semi-Custom Tabulation]." Borealis. https://doi.org/10.5683/SP2/QNO5JG.

von Bergmann, Jens. 2017. "Surprise Maps." https://doodles.mountainmath.ca/posts/2017-04-10-surprise/.

———. 2020a. "Toward Universal TongFen: Change in Polling District Voting Patterns." https://doodles.mountainmath.ca/posts/2020-05-20-toward-universal-tongfen-change-in-polling-district-voting-patterns/.

———. 2020b. "Tongfen: Make Data Based on Different Geographies Comparable." The R Foundation. https://doi.org/10.32614/cran.package.tongfen.

———. 2020c. "TongFen." https://doodles.mountainmath.ca/posts/2020-11-10-tongfen/.

———. 2024. "Population Timelines Update." https://doodles.mountainmath.ca/posts/2024-04-23-population-timelines-update/.