**MGGG-Colorado**

Population Modeling Assessment

-Jose Monge

Current efforts to implement a Gerrymandering-assessment tool in Colorado have been limited by the availability of population data at the appropriate scale, particularly from most recent years. Therefore, population modeling has been used to approximate population distributions for recent years, utilizing custom code and GIS to approximate data at the appropriate scale. Because no approximation is ever perfect, it is imperative to evaluate the in order to understand the limitations of our analysis.

**Population Modeling**

The population modeling was performed under the following premise:

*“The percentage of population from a block-group in a block is (nearly) constant at the two time periods considered”*

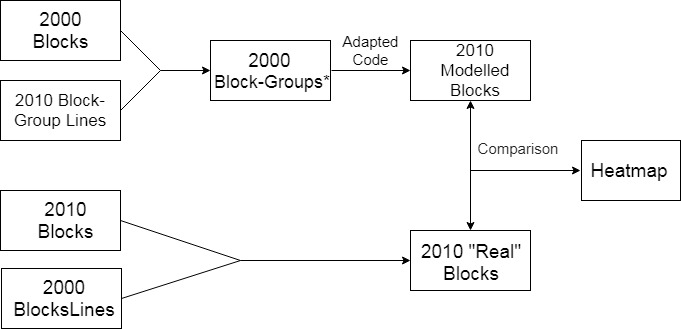
In computational terms, that translates to the following:

As an example, assume that in the year 2000 the population of block-group #10 is of 3000 people, and the population of block #4562 (which is within block-group # 10) is of 300 people. That is, block #4562 contains 10% of the population of block-group #10. Fast forward 10 years, and following migration and population growth changes, the population for block-group #10 in the year 2010 is now of 3100 people. To compute what the “new” population of the block would be, the same 10% is assumed, in which case the “new” population for block #4562 would be of 310 people. Analysis is then carried with that estimated population.

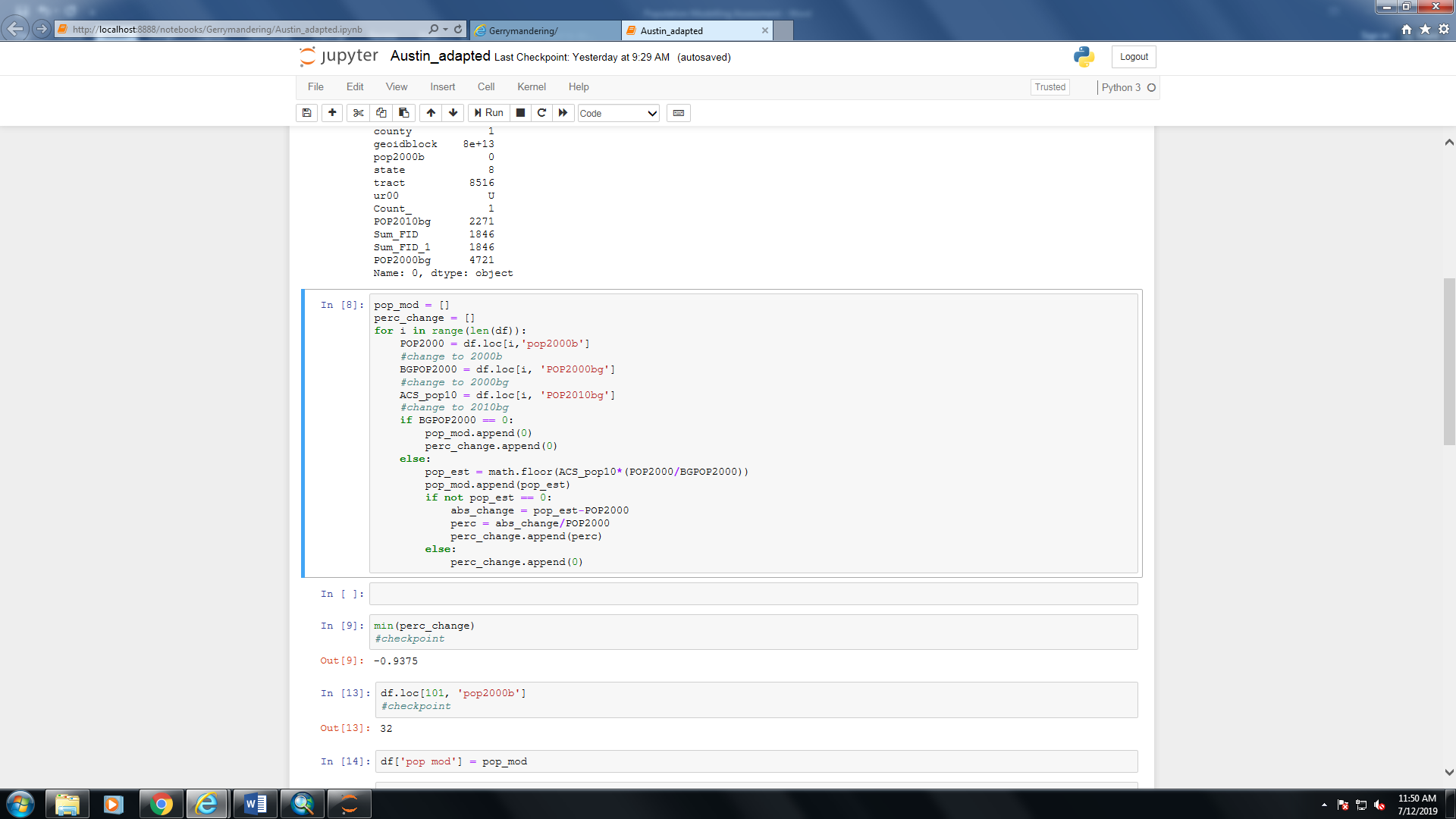
As such, by distributing population according to block groups, a reasonable picture of the population can be obtained, with expected outliers in areas of rapid development such as new housing complexes. The number of blocks (as reported by [State’s Demographer’s Office](https://data.colorado.gov/browse)) in Colorado in the year 2,000 was of 141,040 with 3286 block-groups. By the year 2,010, the same source reports 201,062 blocks and 3,532 block groups (an even the addition of a new county, Broomfield) with a total population growth of 958,000 people, or 23%. The undergoing work of this group implemented the method to take the 2017 population data to project it onto the year 2010, allowing us to compare the results generated by the tool with real recent election results.

**Population Modeling Assessment**

In order to assess the soundness of the population modeling implemented, the 2010 population distribution was compared to what the method estimates the population in that year should have been, streamlined:



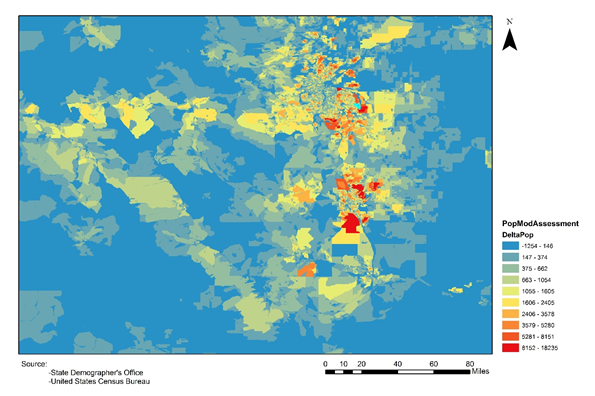
The 2,000 population[[1]](#footnote-2) at the block level was distributing using the boundaries of the 2010 block-groups, then the custom code[[2]](#footnote-3) was run with this input as:



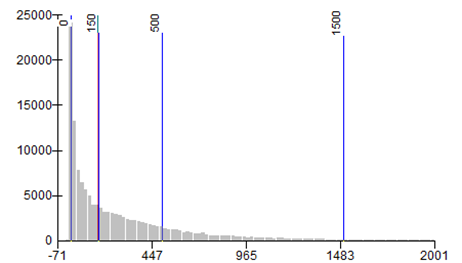
This generated a modelled 2010 population column for each block. This distribution was then compared to the original 2010 population at two scales: block level, and the scale of interest, the precinct level.

**Block Level Comparison**

The following map shows a comparison between the computed and “real” population distributions at the block scale (Min: -1200, Max: 18,000, mode: 0).



The comparison consistently shows the estimated population is smaller than the “real” one (in part due to the floor function used though this was later amended to ceiling function (in some cases the percentage obtained was too small so the round and floor function would project zero from a 1 person block, so ceiling function is argued for. The mode and median are both zero which are positive results, though there are some inconsistent areas which will be discussed below that make this approach somewhat unreliable.



It would be desired that the numbers are packed closely to 0, indicating no large difference in the comparison. Our comparison shows generally encouraging results, as the mode is right at zero and the median is 156. Nonetheless, there are some heavy outliers as well as a long-trailing right skew.

Further analysis indicates that the issues that the comparison seem to point at are not from the modeling itself but from the comparison method. Given that there were vastly different numbers of blocks and block groups in the two time periods used, in order to create a valid comparison, we “redistributed” the 2010 blocks into the 2000 blocks, which created some outlier groupings in which the difference between actual and predicted of the block group is 18,000. As such, the issues elucidated from the map above are not simply from the data but from the comparison model itself, particularly from the grouping. Note that given that the data at the block level is only an intermediate step most of the efforts were put into the precinct level comparison.

**Comparison at the Precinct Level**

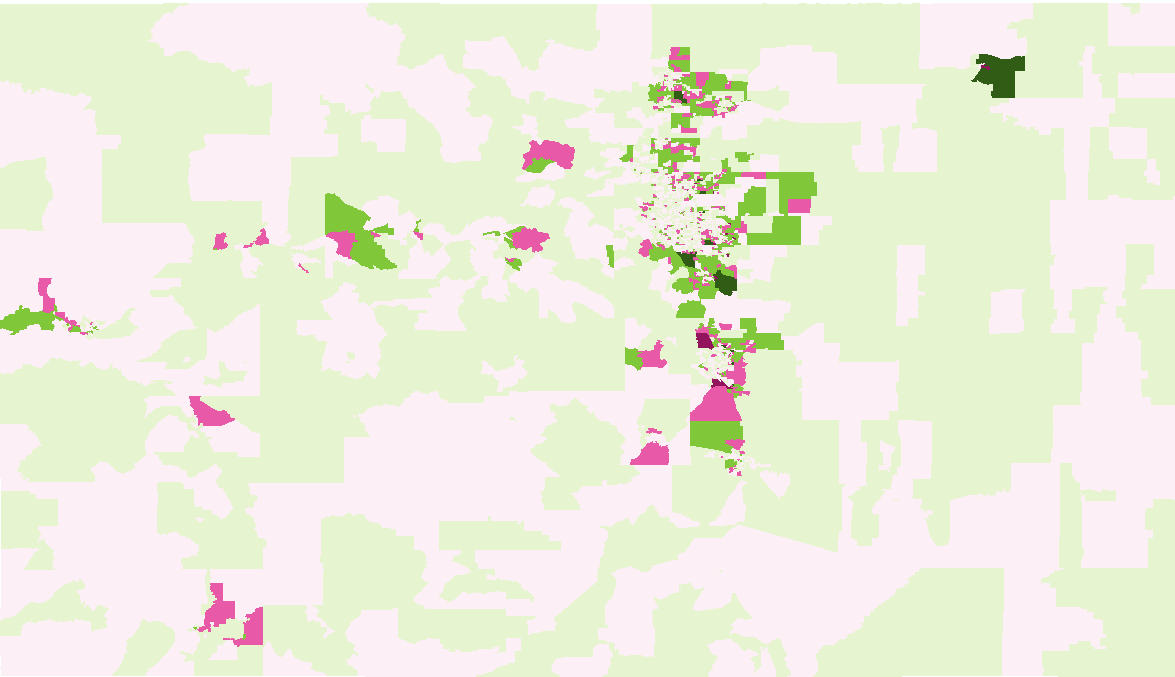
A second method of comparison was implemented to assess the modeling which corresponds to the scale at which GerryChains works. Instead of grouping blocks at the block-group level, blocks were now group into their precincts (which contain many block groups) from the same year (2010). As such, the issues of difference of number of “groupings” was circumvented. The 2010 block data was taken to 2000 block lines, set as centroids (with the shape-to-object ArcMap tool) to subsequently spatially join the most up-to-date precinct lines. The generated 2010 population was then also converted into centroids to spatially join the same precinct lines. The difference between the population in the precincts was the computed as:

Delta\_abs=Real 2010 Population - Estimated 2010 Population

Given the vast differences in population distribution, particularly between the denser urban areas along the I-25 corridor versus the more rural western slope, percentage difference was computed as

delta\_perc = [(Delta\_abs)/(Real 2010 Population)]\*100.

This percent difference was then mapped



It is desired that the percentage difference is as close to zero as possible, but most importantly, absolute number of people being small (note that a large percentage can in fact be simply a small number of people). There are a few cases in which the estimated population was 0 even if the 2000 population wasn’t zero which arose from the floor function but these were a small enough number of cases that it was manually fixed. The other issue is areas were blockgroup 2000 was zero, which would yield a 0/0 case in some instancs (299 blocks). These specific data points were manually identified, and their percentage change was manually set to 100%. Note that these areas were also problematic on the first approach by block (in which non-yet existing blocks were poorly clumped with older blocks, creating large population anomalies, detailed explanation follows below.

**Detailed Methodology**

**Code Changes**

The code used was adapted slightly from the provided version.

1. The switch Low\_memory = False was implemented in the data frame reading for mixed types are present.
2. The math.floor function was changed to round() to avoid losing people to rounding.
3. Instances of Blockgroup2000 being zero and block being zero were changed from append(0) to manually changing both fields to 1 (as to get 1/1, or 100% as 0/0 is not interpreted as 1). Similarly, perc\_change, albeit ignored, was changed to 100%.

**GIS**

Base shapefiles (2000 blocks, 2000 block groups and 2010 block groups) were obtained from <https://data.colorado.gov/browse>. Files referred to in this reference can be found in the folder name Gerrymandering\_Jose\_PopulationModelling. It is recommended to use ArcCatalog to browse for the files mentioned in this section.

**Modelled 2010 Population**

The 2010 Block group shapefile was stripped from its population attributes by the delete fields tool in ArcMap. The resulting file (**lm\_bg.shp**) was then saved separately. The 2000 blocks were turned from polygons into points using the Feature-to-Point tool in ArcMap. Subsequently, these two individual layers were joined using the Spatial Join (Based on Location, SUM) tool in ArcMap (**2000b\_in2010bgl.shp**). The points were added to the block group lines (not the other way around). The SUM function adds the attributes of the points that fall inside of the polygon, in this case it will add the population of the 2000 points that fall inside each 2010 block group. This yielded the 2000 population distributed within the 2010 block groups.

Note that in the table each block is correlated with this corresponding block groups, with many rows repeated for the block groups columns (as many times as there are blocks in each block group). To do this, we did to “reverse joins”, that is spatial joins based on location (same ArcMap tool) from the blocks to each block group. This yielded two maps each with the blocks and their block groups properly indexed. The tables from these two maps were joint (join from table, ArcMap functionality), cleaned and subsequently used as input for the code. Before running the code, the geoidblock column data type was specified as number and the decimal places were removed (in excel itself).

The input (**input\_for\_estimation.csv**) for the population modelling code utilized is a csv table with 141,040 rows and the four following necessary columns:

|  |  |  |  |
| --- | --- | --- | --- |
| GEOIDBLOCK | BlockGroupPopulation2000 | BlockPopulation2000 | BlockGroupPopulation2010 |

The output of the code was another csv file (population\_estimate\_redo5.csv) with two extra columns, PopMod (population modelled) and perc\_change (percentage error) for each block. The attributes of this points are then:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| GEOIDBLOCK | BlockGroupPopulation2000 | BlockPopulation2000 | BlockGroupPopulation2010 | PopMod | Perc\_change |

This table was joint (using join table based on geoidblock ArcMap tool) to the 2000 blocks turned into points layer (**2000b\_barepoints.shp**) to create a final point layer with the modelled population and the columns used to generate the desired shapefile.

**Real 2010 Population**

Given the large difference in number of blocks (as well as geometries), the 2010 population had to be projected into the 2000 blocks. First, a “bare” 2000 block (**2000\_block\_lines.shp**) layer was created by deleting the non-spatial attributes of a fresh 2000 block shapefile, resulting in only block polygons. Then, the 2010 blocks were turned into points (using the Feature to Point ArcMap tool) and spatial join based on location (with SUM) was performed, yielding the 2010 block population distributed onto the 2000 blocks (**2010b\_into\_2000b.shp**). The resulting shapefile had the following attributes:

|  |  |
| --- | --- |
| Geoidblock | Real\_2010\_population |

This table of attributes of tables was the saved independently from the shapefile via export as csv functionality (**real\_2010\_blockpop.csv**).

**Comparison**

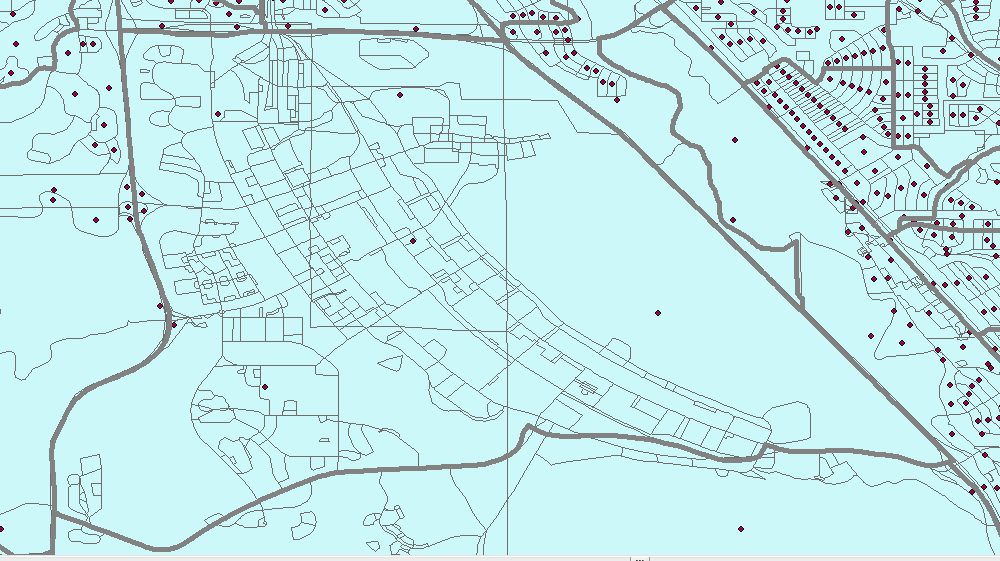
To the attribute table from the modelled points (popmod\_points.shp), the attributes from the real 2010 population in the 2000 block lines (**real\_2010\_blockpop.csv**) was added via join table. Note that geoidblock has been used for indexing, for each block as a unique geoidblock, thus ensuring proper assignation of attributes. The resulting table contains all necessary attributes for the analysis.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Geoidblock | BGpop2000 | BGpop2010 | BPop2000 | Modelled  Population | Real 2010  Population | Difference | % off |

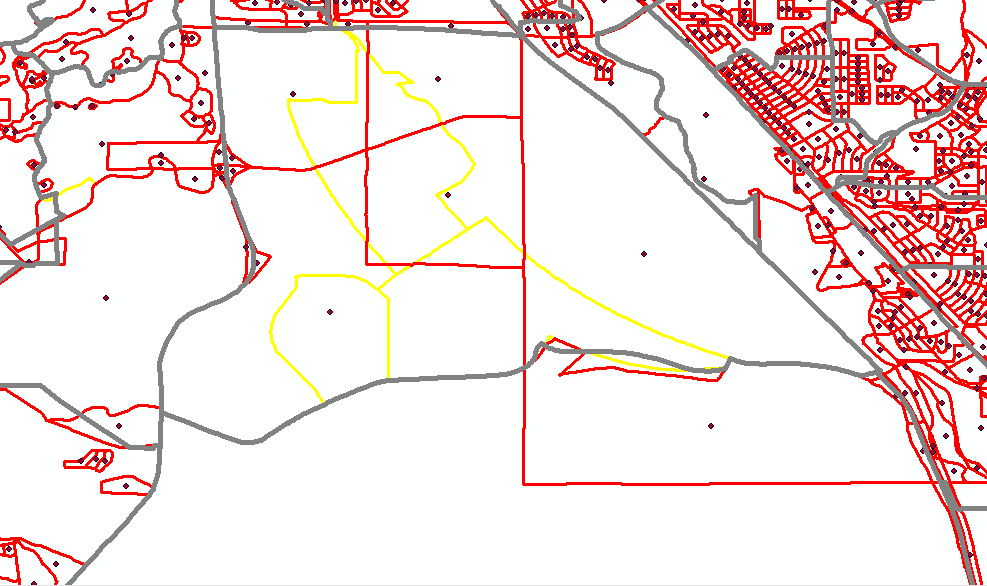
Note that by having these attributes inside of a 2000 block point layer we are able to transform the scale of our analysis to precincts, counties, block groups etc. by simply doing spatial joins of the same nature we used before.

**Issues**

The large incompatibility of the blocks/block groups at the two years did in fact come at a price, particularly in our worst precinct results:



The outlining grey area shows the precinct corresponding to Fort Carson in the year 2010, with the small lines representing the 2010 blocks. Note the points inside are the 2010 population that were turned into the 2000 blocks, which very obviously incorrecct.



In red are highlighted the 2000 blocks, which are of a completely different size as compared to the 2010 analogous. This certainly led to issues, particularly looking at the top-most block which extends far outside of Fort Carson and whose centroid was allocated outside of the block group. Particularly, paying attention at the yellow lines (2010 block groups), it is obvious that the 2000 blocks are not completely functional for there are two block groups that are not even assigned a block (for which we couldn't even then produce a modeled population).

**Additional Adjustments**

A change to table used as input for the code was manually done to improve cases in which the modeled would predict zero population. This would happen in places were the block group population was zero (in the year 2000), thereby automatically assigning a 0% to all blocks contained within the block group. This would cause that any block group that in 2000 had zero population but in 2010 had some population (as seen in instances of large developments and military installations) would be multiplied by zero, thereby making the modelized population highly inaccurate in those specific block groups. To counter-act this, to all the block groups with population zero, we manually changed the population to 1, and to each of the blocks contained within the block group we assigned a fraction equaling 1 divided by the number of blocks in that blockgroup. For example, think of a block group with population zero and 3 blocks. The block group population is changed to one, and to each block a 1/3 is assigned, thereby artificially modelling population for that block instead of predicting 0. Note that the data type for BPOP2000 was changed from int to float (as float(BPOP2000) in the code).

There are also instances in which the percentage is so small that the floor function makes the pop\_mod zero instead of one. Given that there are many blocks with very small population, this issue could have an impact of around 141,000 underprediction on the model. Round function (from math.floor to round()) or math.ceiling are suggested instead.

The column perc\_change produced by the code was not considered for the assessment. This is because of the significance of the numbers it shows. A very large percentage discrepancy (6000%) could be a mere 60 people off, but a small small percentage in a population-heavy block can be way larger (10% of 5000 which is 500 people). For the analysis, the absolute number of people is a more important metric. Therefore, three new statistical fields were added. First, a simple Difference= “Real \_2010\_population” - “Modelled\_2010\_population” which displays where the model’s absolute discrepancies are larger. This field but in its absolute value form was also produced, preventing from negatives and positves from cancelling and falsely improving our results. The third statistically relevant column is discrepancy divided by state population. This gives an absolute “ruler” for all the discrepancies across the state.

**Pending Issues**

The main issue is the instances in which invidual blocks from a block group added together do not add up to the block group population. This is observed in the “outlier” regions though it could be a symptom of an issue that is not as notable elsewhere. Kading Mangalik is currently continuing this work to identify the cause and solution to this issue.

**Latest**

After not being able to successfully resolve the issues arising from this specific approach, it was decided that given that the precinct lines and block group lines are anyway not completely compatible, the 2000 block group lines would be used instead, thereby alleviating the issues that arouse from the incompatibility with the 2010 blocks.

1. Which is at the best possible level of accuracy as it is reported by the census. [↑](#footnote-ref-2)
2. Built by Austin Eide, B.A. [↑](#footnote-ref-3)