10 August 2023

Carola,

We are hereby resubmitting our manuscript, "Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites", to the Research Spotlights section of *SIAM Review*. In his communication of the referee reports, Des Higham mentioned that you would be handling the paper by our planned resubmission timetable.

We thank the referees for their positive feedback and their helpful suggestions. We highlight the key changes in the revised manuscript document, and we give point-by-point responses to their comments below. We have also carefully gone through the manuscript and made our own expository improvements throughout.

We look forward to receiving feedback on the revised manuscript.

Sincerely,

Mason (on behalf of all authors)

Referee #1 (Remarks to the Author):

Overall this is a good paper with an interesting and practical application of persistent homology. Descriptions are clear and precise. The authors are thoughtful with how they define a metric, relying on travel times and wait times over geographic distance. Their models incorporate on multiple data sources for multiple cities and their interpretation was meaningful. This seems to be an improvement upon other related works. They do address that given more fine-grain data, their model would improve in accuracy.

There are a few questions that I would like to see briefly addressed.

1. Why isn't the distance function symmetric if you are including travel time from x to y and back to x again? Naively, I would think this is symmetric. Could you give a little intuition as to why it is not? And how does the weighted average address this? I'm not quite following why the weighted average is needed.

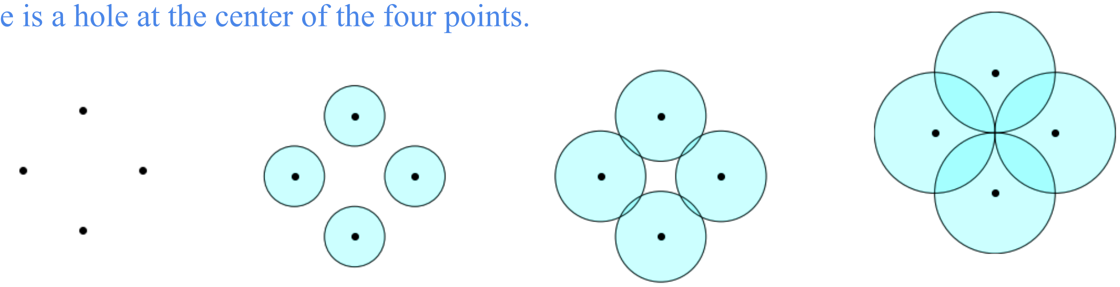
Response: The asymmetry is due to the car ownership *C(Z(x))* term in the definition of \tilde{*d*}*(x, y)*. Even though the *t*_{car}, *t*_{pub}, and *t*_{walk} matrices are symmetric, it is not necessarily the case that *C(Z(x))* = *C(Z(y)*), so in turn it need not be true that \tilde{*d*}*(x, y)* = \tilde{*d*}*(y, x)*. We have added a comment in the paper to emphasize this point (see lines 243–244).

2. In the results section, can you comment on the difference between what 0-D and 1-D homology tells us? The interpretation of holes in coverage captured by H_1 is clear, but I would like a little more about H_0 and why you include it in your discussion. (I do think it's helpful, but should be discussed more explicitly).

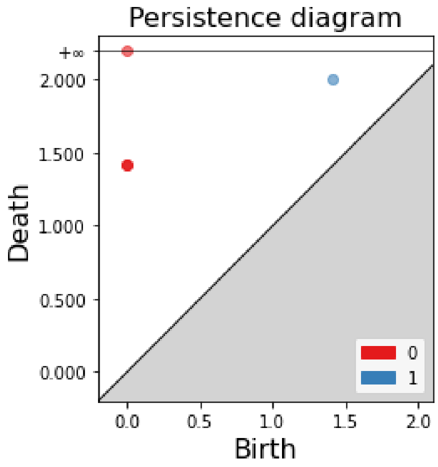
Response: We have added a short discussion in lines 323–326.

3. Should I really be thinking of the very small lifetimes, but high death values of H_1 in Salt Lake City as significant holes in coverage?

Response: In the filtration {*B(x, r_x(t))*}_{*t* \in *R*} of surfaces (i.e., 2D subsets of the plane), a 1D homology class with a short lifetime but large death value can still be a significant hole in coverage. We show a toy example here. Even though the 1D homology class has a relatively short lifetime, its large death value correctly reflects the fact that there is a hole at the center of the four points.



Below is the PD for the toy example above.



This example aside, the reviewer is correct to be concerned that homology classes with extremely short lifetimes, like those in Salt Lake City, may be artifacts of our Vietoris–Rips approximation. To account for this possibility, we have modified our analysis to exclude homology classes whose lifetimes are not sufficiently long. In our new analysis—specifically, in Table 1 (median/variance of homology-class death values), Figure 6 (box plot of the death-value distributions), Figure 7 (a comparison of the death-value distributions in Atlanta and Chicago), Figure 8 (death simplices for the 0D homology classes), and Figure 9 (death simplices for the 1D homology classes)—we consider only homology classes whose death/birth ratios are at least 1.05. In the PDs, we still show all homology classes. We have added a discussion about short-lifetime homology classes to the paragraph on lines 331-339.

The death simplices shown in Fig 9f seem to span multiple precincts. Can you comment briefly on why the death simplices in Fig 9 often overlap other polling sites? Is this an artifact of the distance used?

Response: Our "metric"[1] is not Euclidean, but in Figure 9, we plotted the death simplices as Euclidean triangles. The Euclidean triangles sometimes include other polling sites. However, the geodesic triangles may not contain those polling sites. It is computationally infeasible to plot the geodesic triangles because it would require computing shortest paths (with respect to our metric) between each pair of points in the triangle.This would require many more travel-time queries, which have a significant monetary cost. We believe that the overlaps with other polling sites are primarily the result of our non-Euclidean metric.

Another possibility is that a polling site could have such a long waiting time that it does not show up in the filtration until after the homology class (the class whose death simplex overlaps the polling site) has already died.

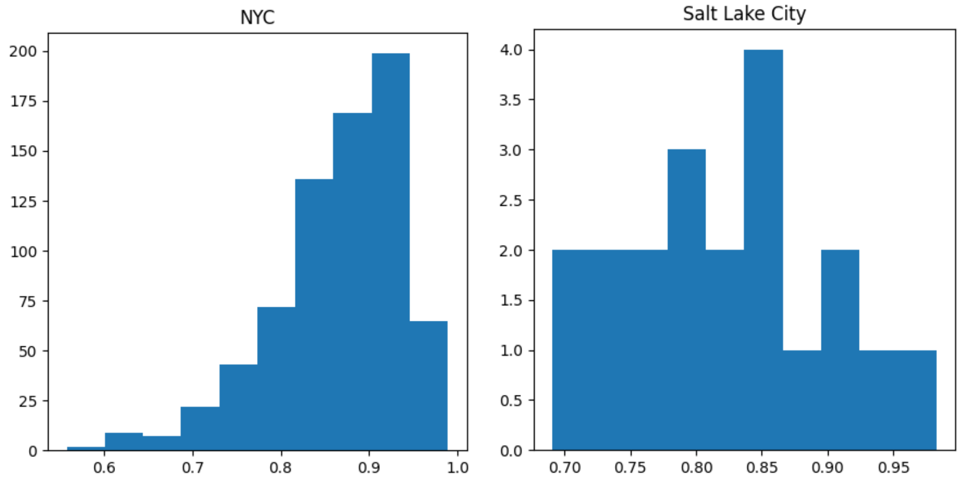
We have added a footnote (line 359) to discuss this point.

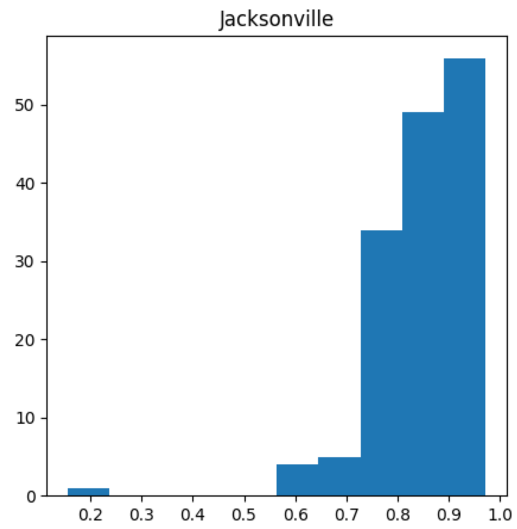
4. I am a little concerned that the convexity condition of the Nerve theorem is not met. I know you address this briefly in the limitations section, but can you mention why you do not think its a problem in this application? Do you think it is creating artificial holes?

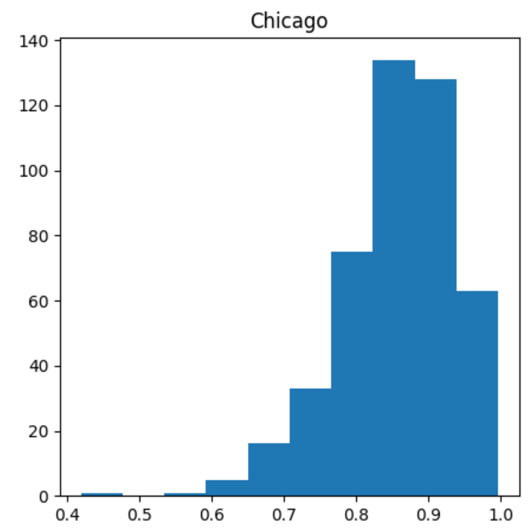
[1] Technically, the function *d* that we define is not a metric because it does not satisfy the triangle inequality.

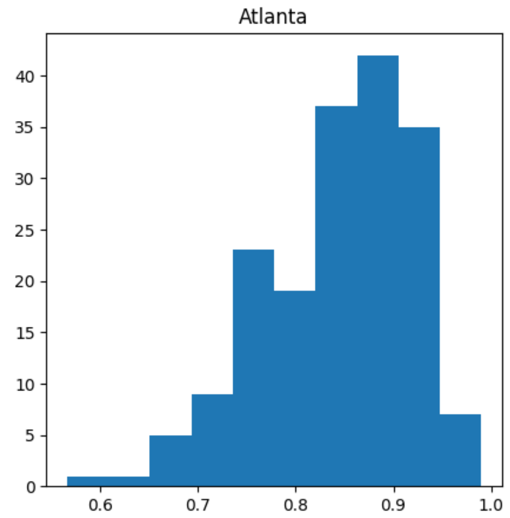
Response: The convexity condition of the Nerve Theorem is not met because our metric is not Euclidean. However, it is still reasonable to assume that our metric is approximately *locally* Euclidean. This is the case because car-ownership rates and traffic conditions do not vary much within a small neighborhood. Because our metric is locally Euclidean, sufficiently small balls (with respect to our metric) behave like Euclidean balls, so the Nerve Theorem is applicable for sufficiently small filtration values. We have added this explanation to a footnote on line 422.

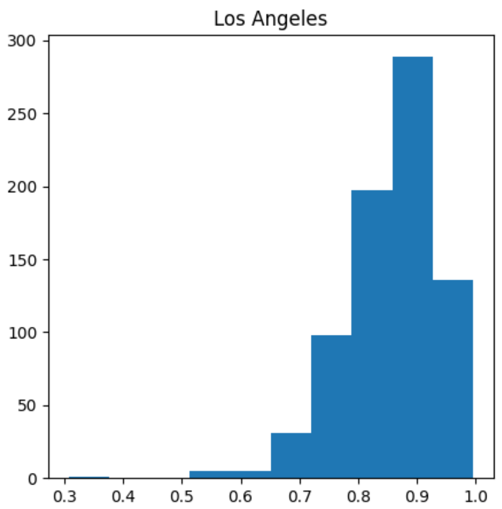
We verified the above claim empirically as follows: For each polling site x, we calculated its $k =$ 15 nearest polling sites with respect to our metric d. We then calculated $d(x, y)$ and $d\_E(x, y)$ for each nearest neighbor y, where d_E is the Euclidean metric. We computed the Pearson correlation coefficient between the set $\{d(x, y) \mid y$ is one of the 15 nearest neighbors of $x\}$ of distances measured using our metric and the set $\{d(x, y) \mid y$ is one of the 15 nearest neighbors of $x\}$ of Euclidean distances. Below, we show the histograms of Pearson correlation coefficients for each polling site x in each city.











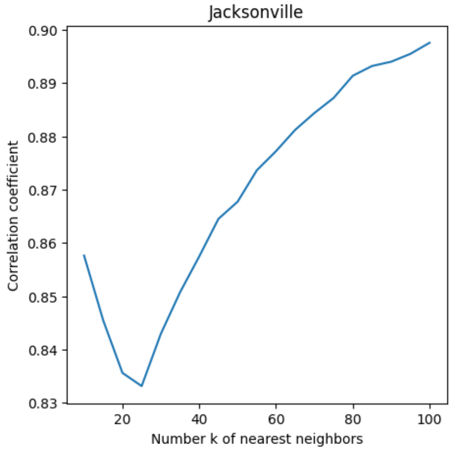
For each city (with Los Angeles technically referring to the whole county, as discussed in the manuscript), here are the mean Pearson correlation coefficients:

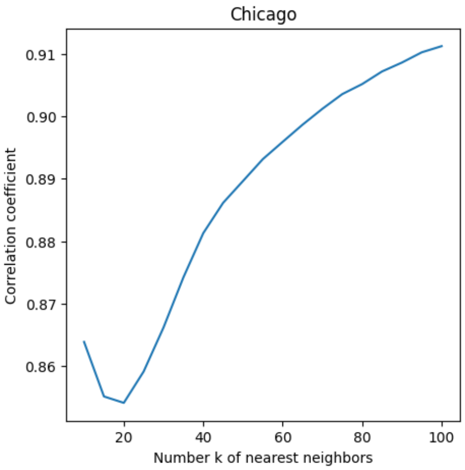
City	Mean correlation coefficient
NYC	0.8656
Salt Lake City	0.8231
Jacksonville	0.8454
Chicago	0.8551
Atlanta	0.8422

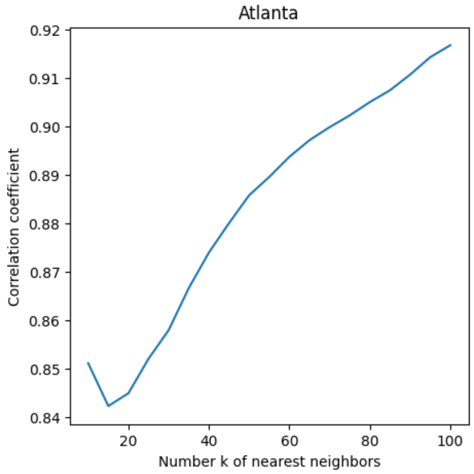
Los Angeles County	0.8569

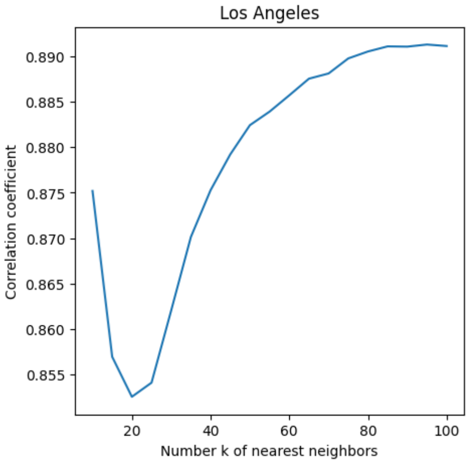
These correlation coefficients are between 0.82 and 0.87, which implies a strong linear correlation. This indicates that if *x* is a polling site and *y* is sufficiently close (i.e., within the 15 nearest neighbors of *x*), then d(x, y) \approx a*d_E(x, y) + b, where *a* and *b* are constants that depend on *x*. That is, our metric is locally Euclidean. (Note, however, that our metric is not *globally* Euclidean because the constants *a* and *b* depend on the polling site.)

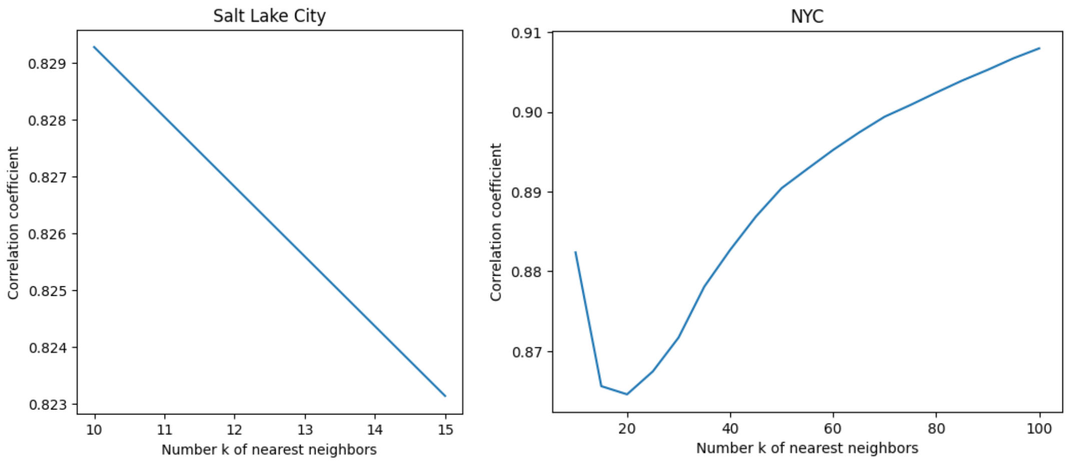
How local is “locally Euclidean”? For each city, we performed the same experiment as above with *k* = 10, 15, …, min{100, number of polling sites in the city} nearest neighbors:











These correlation coefficients are very large even for large k, so our metric is approximately locally Euclidean for all examined values of k.The balls that we analyzed above have radii on par with the length scale of interest. In the table below, we report (for each city and LA county) the mean distance (with respect to our metric) to the $k = \min\{100,$ number of polling sites in the city$\}$ nearest neighbors:

City	Mean distance (minutes)
Atlanta	104.7977
Chicago	81.8094
Jacksonville	118.48078
Los Angeles County	105.4005
Salt Lake City	59.5395
NYC	100.1955

We conclude that it is reasonable to assume that our balls are “approximately” convex for the filtration parameters in the range of interest.

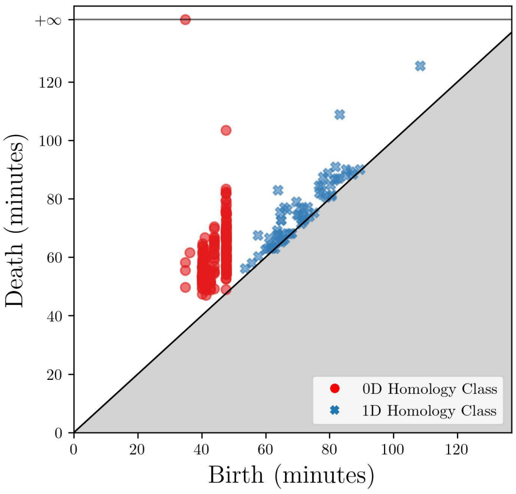
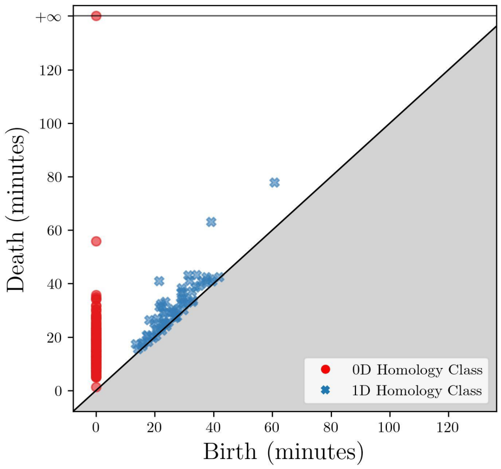
Referee #2 (Remarks to the Author):

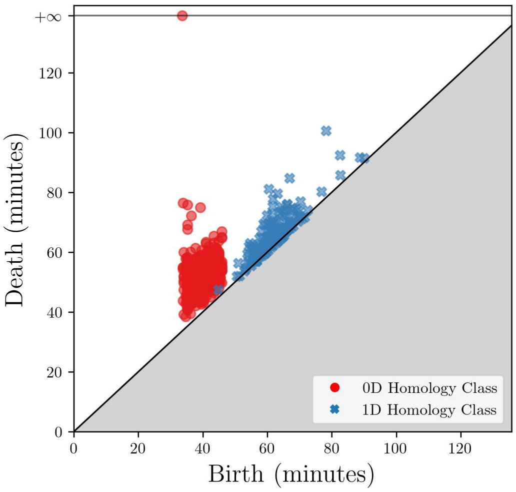
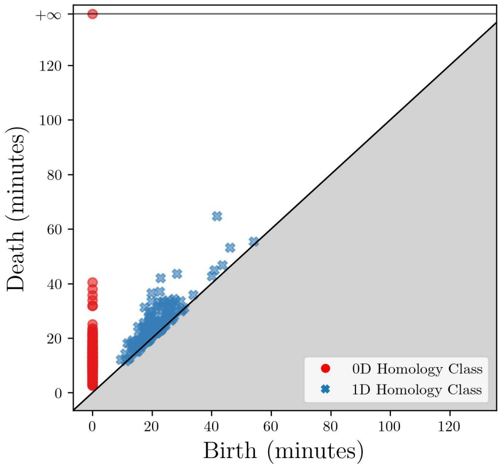
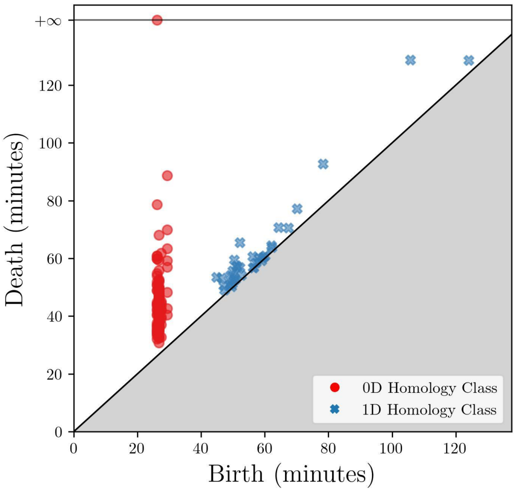
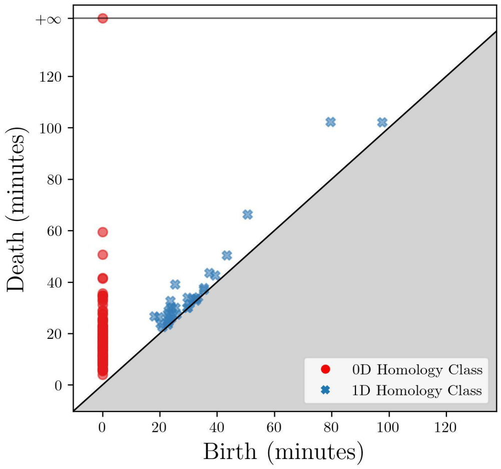
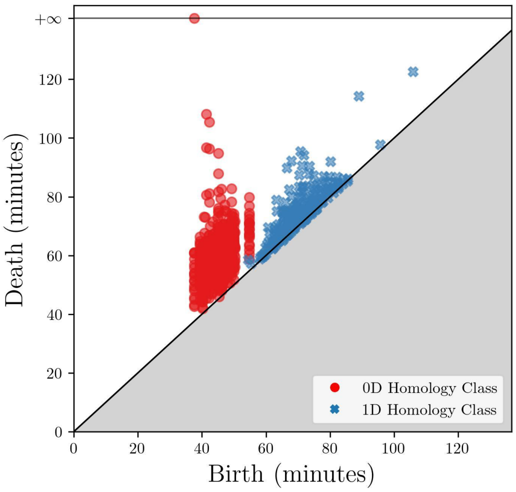
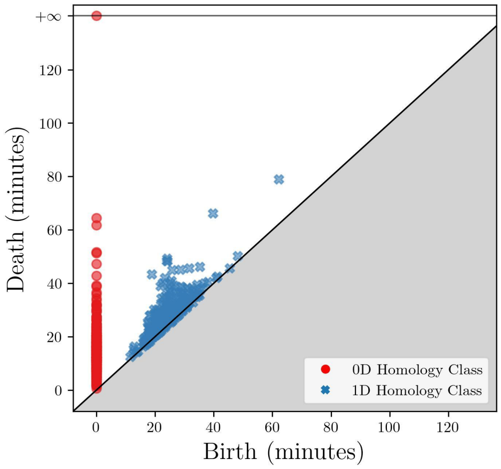
The use of TDA and PH to study access to polling sites appears to be a new and interesting idea. The analysis performed by the authors is correct and meaningful, and the results are intriguing, despite not representing a breakthrough.

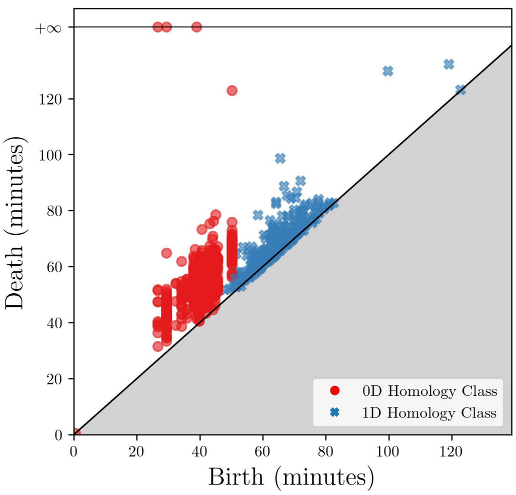
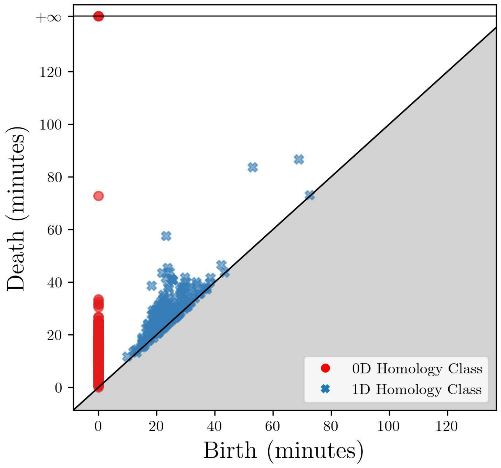
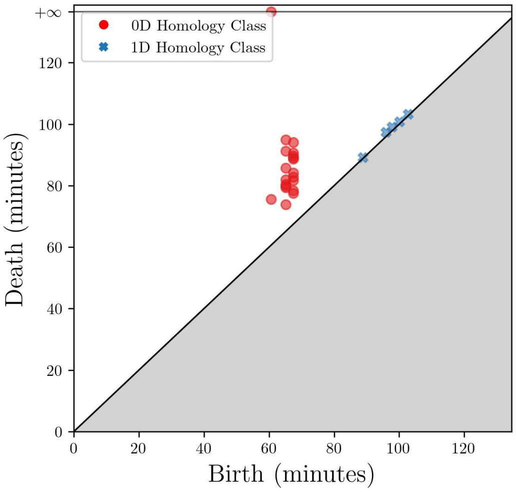
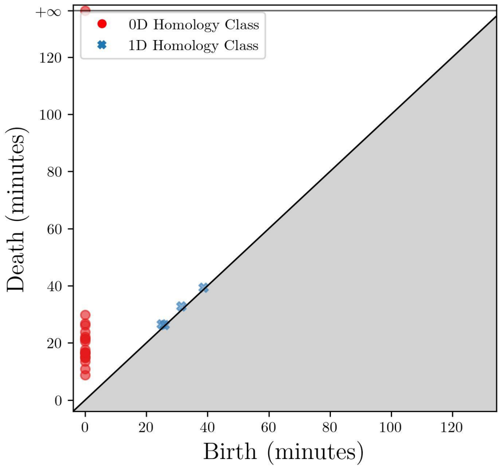
Some specific comment that I think might improve the manuscript, and that I would suggest the authors to address.

- The authors construct a weighted VR filtration; although this looks like a reasonable choice, and even if the specific choice of weights is well motivated, I wonder whether it is necessary. In other words, how different would the results be disregarding weights?

Response: We recomputed the persistent homology for each city (and Los Angeles County) using a VR filtration based only on travel times. We show our results below. Even though our waiting-time data is very coarse, we still observe that the PDs are visually very different when we remove the weights. We expect that there would have been an even greater difference had we had granular waiting-time data for each polling site.

	With weights	Without Weights
Atlanta		

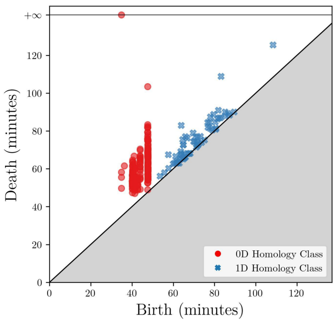
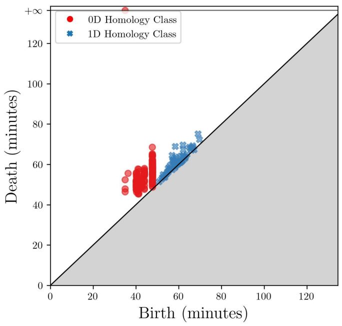
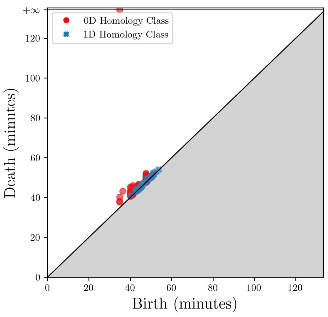
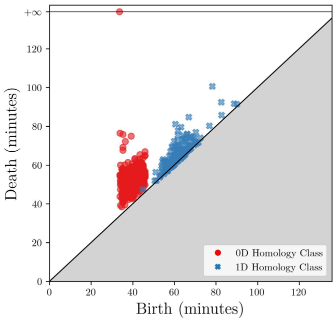
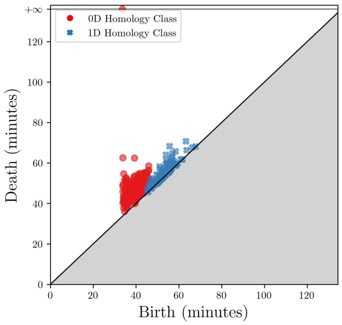
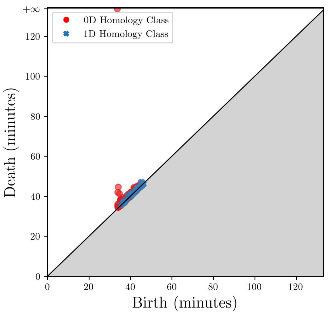
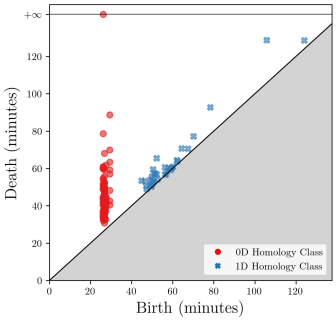
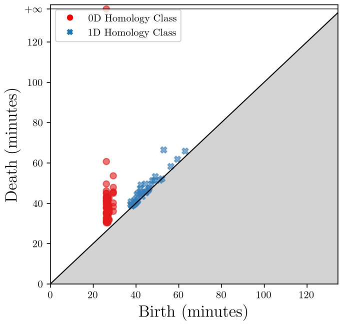
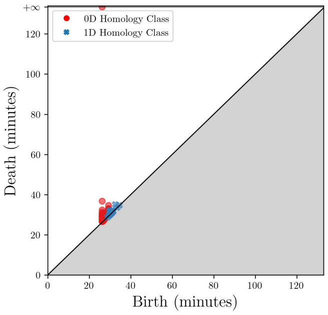
Chicago		
Jacksonville		
Los Angeles County		

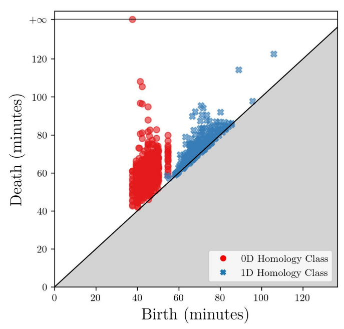
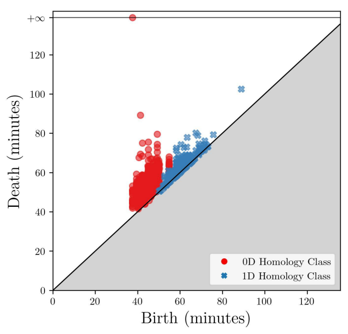
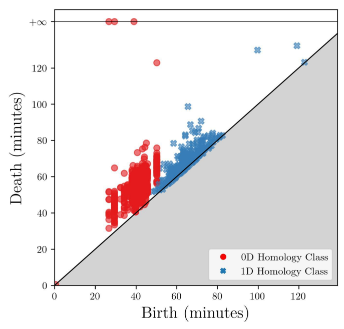
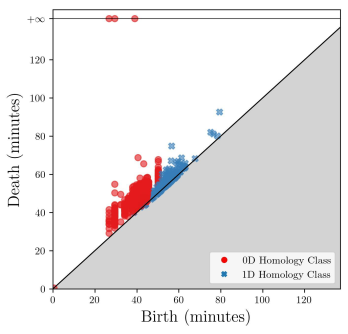
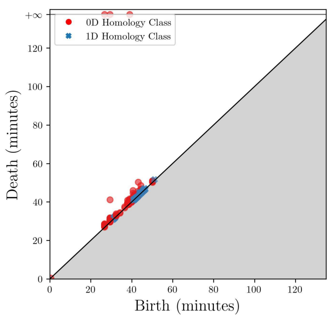
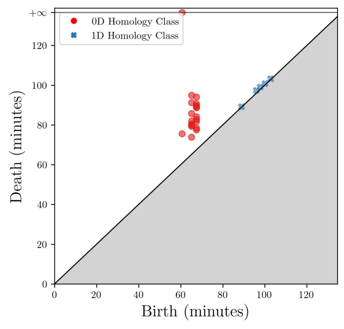
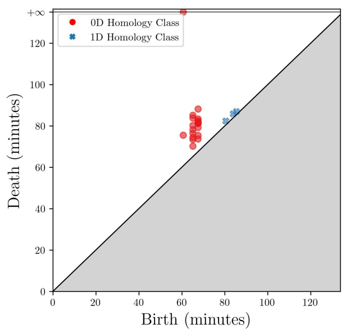
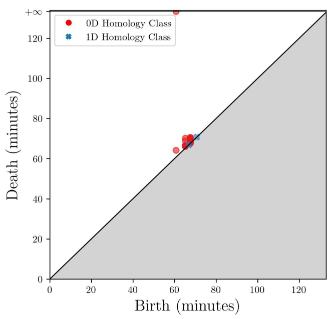
New York City		
Salt Lake City		

- Similarly, have the authors considered varying their metric of choice, and looking at how the PDs change if restricting to only one or two means of transport (e.g. what happens if we exclude car or public transport)?

Response: We recomputed the persistent homology for each city (and Los Angeles County) using a "metric" (note that it is not technically a metric, as it does not satisfy the triangle inequality) that excludes public-transportation time. We assumed that everyone has access to a car and defined *d(x, y)* to be the travel time by car from *x* to *y* and back. We also recomputed persistent homology with the additional modification that we estimated car travel times using OpenStreetMaps (OSM) instead of using Google data. We assumed that cars travel at the speed limit along each road. We show our results below. For each city, these recomputed PDs (which we label as “Without public transportation” and “Without public transportation and with OSM”) are visually very different from the one in the paper (which we label as “With public transportation”).

In LA, we were unable to compute PH using OSM because we had trouble computing some of the distances. However, based on the results for the other cities, we would be very surprised if the PD for LA with OSM did not differ from the other PDs that we computed for LA.

<table>
<tr><th></th><th>With public transportation</th><th>Without public transportation</th><th>Without public transportation and with OSM</th></tr>
<tr><td>Atlanta</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Chicago</td><td>
</td><td>
</td><td>
</td></tr>
<tr><td>Jacksonville</td><td>
</td><td>
</td><td>
</td></tr>
</table>

Los Angeles County			n/a
NYC			
Salt Lake City			

- The past few years have seen an important improvement in TDA's interpretability power, and there are now several ways to locate topological features in the data (e.g. by using generators https://www.frontiersin.org/articles/10.3389/frai.2021.681117/full; https://arxiv.org/abs/2210.07545; or decorated merge trees https://arxiv.org/abs/2103.15804). I think that this would potentially allow to analyse how accessibility varies for different areas in a city, and perhaps try to relate that to racial/economics demographics, substantially increasing the interest of the study and results.

Response: This is an interesting question that is worthy of investigation. Therefore, we have added a short discussion of these ideas to our "future work" section. (See lines 455–463.) However, pursuing these ideas is beyond the scope of the current paper.

One can use minimal generators to calculate representative cycles with minimal length. Each cycle would encircle a hole in coverage, allowing one to find the boundary of a hole, whereas death simplices allow one to locate the “epicenters” of the holes in coverage.

Given minimal generators, one could use “hyperTDA” (see https://arxiv.org/abs/2210.07545) to analyze the structure of the minimal generators. In hyperTDA, one builds a hypergraph in which each vertex is a point (i.e., polling site) in a point cloud and a hyperedge connects the vertices in a minimal generator. Then, as described in the hyperTDA paper, one can use hypergraph centrality measures and community-detection methods to analyze the minimal generators. This method might provide insights about the spatial structure of the minimal generators.

Decorated merge trees (DMTs) are another potential approach to use to locate the 1D holes in coverage. DMTs allow one to identify homological cycles with the cluster of points (in the point cloud) in which they belong.