

QUANTIFYING NOTHING

Caterina Chiopris*

July 2026

Abstract. Many questions in social science aim at capturing a lack of resources, the missingness of people and groups, or gaps in representation. Yet, most methods are developed to measure positive or negative quantities, rather than the size of absences. We introduce ‘nothing’ as a quantity of interest in and of itself, as a bounded absence enclosed by a presence, with a size, a location and a scale, and develop desiderata for its measurement. We show how thinking of data as shapes, and using topology applied to data analysis, provides a tool to uncover the existence and relative sizes of holes, or ‘nothings’, that is robust to perturbations and to continuous deformations of the space. We then show how our method can detect food deserts without relying on a given threshold or scale, how it identifies the effects of genocide as missing generations, and how it can detect voids in representation based on voters’ and parties’ ideal points. We thus assign a measure to what previously could only be seen.

Keywords: measurement, topological data analysis, persistent homology, resource gaps, spatial analysis

*Columbia University, cc5398@columbia.edu

I am grateful to James Chu, Don Green and Gary King for suggestions and early encouragement, and to the Columbia Data Science Institute’s Computational Social Science Working Group for comments. Haney Maxwell provided the most excellent research assistance.

1 Introduction

Within empirical political science, and the social sciences more broadly, we normally think of measurement as quantifying the value of a variable that exists, and that could be positive or negative. We construct measures of polarization, literacy, or ethnic fragmentation that run from 0 to 1; we measure tax rates and incomes, count word frequencies, and assign sentiment to a speech through a number. This presupposes the *existence* of something: taxes, income, sentiment, ethnicity, to which we assign a value we can interpret. But the answers to many questions in social science rely instead on quantifying the *absence* of something. Where are resources *missing*? Who is *lacking* representation? Where is there a gap in information transmission? Normally, we do not think systematically about measuring the lack of a variable, or the absence of things. How do you know something is missing? How do you quantify the size of nothing?

The challenge is not that absences are rare, it is that our tools are built for presence. Consider missing polling stations, or a district without minorities, a bloc of voters without representation: each is a ‘hole’ in a distribution, it is in itself substantively an object of interest, yet it is normally studied by measuring something else and reading the absence off indirectly, such as an average that runs low, a distance that runs high, or a count that hits zero. We think instead of an absence as an entity with a location and a size of its own: we note that one large gap is not the same as several small ones, and whether a region looks empty at all depends on how closely one looks. This cannot directly be captured by the mean of a variable defined on pre-drawn units, or with distances with given thresholds.

The goal of this paper is to introduce “nothing” as a quantity of interest in its own right, and to propose a method, adapted from topology, to measure it: to detect gaps or holes, place them in space, and size them against one another. We first define the estimand, an absence enclosed by an existence, that comes with a location, and a size, independently of any procedure for recovering it (Section 2). We then propose a measure (Section 3). The key step is to treat the data not as a list of numbers to be summarized, but as a *shape*, and to borrow from topology, the branch of mathematics devoted to shape, a tool called *persistent homology*, that allows us to count a shape’s holes, and to measure how pronounced each one is. The appeal thereof is that it reads a gap directly from the arrangement of the data, without being told where a gap is, and it does so at every scale at once, so that the answer does not turn on an arbitrary choice of threshold or unit.

We then show how the method can be used, and what the output would be, in three illustrative settings (Section 4). First, we identify food deserts from the spatial distribution of the population receiving SNAP benefits, measuring where poor households are both dense

and far from a grocer, and how severe each such pocket is. Second, we show that the Khmer Rouge genocide left an unusual gap in the male population of Cambodia, and that the deficit is legible, decades later, in exactly the birth cohorts and the districts the historical record predicts. Third, we map which parts of a two-dimensional ideological space are represented by no party in European democracies. In each case the quantity we recover, the located, sized absence, is one that, to our knowledge, no existing method measures directly and systematically, and in each case we choose a setting where the answer is known in outline or is clearly visible, so that the method can be validated against something we can already see.

A recurring theme is that many questions phrased as being about a *something* have a more revealing formulation as being about a *nothing*. Rather than measuring the level of a variable, one can measure the absence of its opposite: not how well the average voter is represented, but where the largest representation gap is; not the mean distance to a grocer, but the deepest enclosed pocket that has none. The two are not the same computation, and the second does not follow from the first. This reframing is familiar in the discipline under other names, the exercise of power through non-decision-making and issues being kept off the agenda (Bachrach and Baratz, 1962), the count of missing answers from surveys, or the leverage of the negative case, such as the war that did not start, the movement that never formed (Mahoney and Goertz, 2004), but it has lacked a general measurement tool. This paper aims at offering one, and it hopes not to be the last.

The paper speaks to three literatures. The first is measurement. A long tradition distinguishes a concept from its operationalization and asks how faithfully the latter captures the former (Adcock and Collier, 2001), with a smaller strand tackling concepts that researchers recognize on sight but struggle to quantify exactly, such as the compactness of a legislative district (Kaufman, King and Komisarchik, 2021). Absence is such a concept: everyone can point to a food desert or an unrepresented voter on a map, yet the field has no general way to quantify one. We contribute the estimand itself, absence as a located, sized object, and a set of criteria a measure of it should meet.

The second is topological data analysis (TDA). Persistent homology grew out of work on computing and simplifying the topology of data (Edelsbrunner, Letscher and Zomorodian, 2002; Zomorodian and Carlsson, 2004; Carlsson, 2009), with a related lineage in the Mapper algorithm (Singh, Mémoli and Carlsson, 2007; Lum et al., 2013). In the natural sciences it has been used in a variety of applications, such as to identify a subgroup of breast cancers with a distinctive survival profile (Nicolau, Levine and Carlsson, 2011), to characterize protein structure and folding (Xia and Wei, 2014), to reconstruct viral evolution (Chan, Carlsson and Rabadán, 2013), and to map the filamentary structure of the cosmic web (Sousbie, 2011); surveys now exist for biology (Amézquita et al., 2020), biomedicine (Skaf and Laubenbacher,

2022), networks (Aktas, Akbas and El Fatmaoui, 2019) and others. Its footprint in social science is far smaller (Feng and Porter, 2021; Duchin, Needham and Weighill, 2022; Hickok et al., 2024). We depart from this work in framing and in scope. Rather than a single application, we propose absence as a general quantity of interest and persistent homology as a general template for measuring it, we handle the common case in which a gap is defined by two conditions at once, population present *and* resource absent, through a two-parameter sweep, and we carry the same construction across three substantively unrelated domains to show that the template travels.

The third contribution is to the substantive literatures the applications touch. For food deserts, three decades of research still seem to lack a unifying framework because findings move with the chosen buffer distance, store definition, and areal unit, a version of the modifiable areal unit problem (Lytle, 2009; Apparicio, Cloutier and Shearmur, 2007; Beaulac, Kristjansson and Cummins, 2009); framing the desert as a scale-free hole sidesteps the arbitrary cutoff, and we compare our results against the canonical racial-disparity finding of Zenk et al. (2005). For genocide, a demographic literature has documented the sex-selective mortality of the Khmer Rouge years and the “demographic scar” its cohorts still carry (de Walque, 2005; Heuveline and Poch, 2007; Gaikwad, Lin and Zucker, 2023); we add a way to locate that scar in space and to size it, corroborated against an independent record of documented mass graves. For representation, congruence is conventionally thought of as an average citizen–party distance (Golder and Stramski, 2010), which reports how well the typical voter is served but not which regions of preference space are served by no one, nor how large those regions are (Lefkofridi, Wagner and Willmann, 2014). We recover the unrepresented regions as objects, and are able to rank them, and we show visually that the different ‘shapes’ underrepresentation can have would make it hard to be fully detected otherwise.

The remainder of the paper proceeds as follows. Section 2 defines nothing as a quantity of interest and states the criteria a measure of it should satisfy. Section 3 develops the method, building from data-as-shapes to persistent homology, and illustrates it by showing it can find Central Park as a large gap in Manhattan’s population. Section 4 presents the three applications. Section 5 concludes, and points to where further methodological work is needed, as well as what the possible applications of the method could be.

2 Nothing as a Quantity of Interest

Definition attempt. What is ‘nothing’? An immediate answer could be: nothing is wherever something is zero. More formally, let the something live on a space $X \subseteq \mathbb{R}^d$, and let $f : X \rightarrow [0, \infty)$ record how much of it there is at each point: grocery stores per square kilometer, voters per unit of ideological space. The tempting definition is then the *zero set* of f ,

$$N_0 := f^{-1}(0) = \{x \in X : f(x) = 0\}, \quad (1)$$

A ‘nothing’ would then be each of the contiguous pieces of this zero set: decompose N_0 into its maximal connected pieces, $N_0 = \bigcup_k C_k$, and report each piece together with its size $\lambda(C_k)$, where λ is the ordinary measure of volume. In one dimension, this is simply a list of the intervals on which f is 0, each measured by its length. Note that dimension is not the obstacle here: the same carries over to any dimension d , with intervals of zeros replaced by zero-connected-components and lengths by areas or volumes. The practical issues with this definition lie elsewhere, and there are (at least) two of them.

The Resolution. Definition (1) presumes we observe the function f . We do not; we observe data. Suppose presence is recorded at finitely many locations $D = \{x_1, \dots, x_n\} \subset X$, the addresses of the grocers, the positions of the parties. Taken literally, the observed f is zero everywhere off the sample, so that $N_0 = X \setminus D$ and $\lambda(N_0) = \lambda(X)$: almost everything is nothing, at every point except the n we observed. To speak of gaps at all, presence must be granted some reach beyond the bare observation, and any such grant involves a scale. This corresponds to assuming that a grocery store does not serve only the space within its physical walls, but some area around it, say, within two miles of its location. To capture this, fix a radius $r > 0$ and declare presence to extend distance r around each observation:

$$P_r := \bigcup_{i=1}^n B(x_i, r), \quad N_r := X \setminus P_r, \quad (2)$$

where $B(x, r)$ is the ball of radius r centered at x .¹ The nothing set now depends on r , and the dependence does matter. To see why, consider the smallest gap between any two observations. Then, halve it, and call it $\underline{r} := \frac{1}{2} \min_{i \neq j} \|x_i - x_j\|$. Similarly, consider the farthest any location sits from its nearest observation, and call it $\bar{r} := \max_{x \in X} \min_i \|x - x_i\|$.

¹When the data are not a list of locations but a measured field f on the space, as in our applications below, the same role is played by a threshold: $P_\tau := \{x : f(x) \geq \tau\}$, with the sweep running over τ instead of r . The zero set (1) is the single slice $\tau \rightarrow 0$ of that family. Everything in this section applies with τ in place of r .

Then for every finite dataset, whatever the underlying reality:

- if $r < \underline{r}$, the balls in (2) are pairwise disjoint: every observation is an isolated island, and every space between neighboring observations is a gap;
- if $r > \bar{r}$, then $P_r \supseteq X$ and $N_r = \emptyset$: everything is connected, and there are no gaps at all.

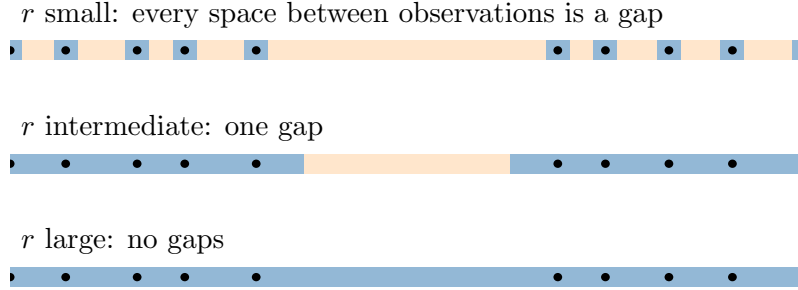


Figure 1: The same ten observations (dots) at three scales. Presence at scale r (blue) extends distance r around each observation; what remains uncovered (beige) is the nothing set N_r . At a fine enough scale every observation is an island and gaps are everywhere; at a coarse enough scale no gap survives.

Both extremes are degenerate: at a fine enough scale there are always holes between observations, and at a coarse enough scale there are never any. Figure 1 provides a visual intuition for this. It shows the same ten observations at three scales, holding one, many, or no gaps. In between the two extremes the family is at least orderly, $r \leq r'$ implies $N_{r'} \subseteq N_r$, so gaps only ever shrink and vanish as the scale coarsens; but nothing in the data singles out a correct r at which to stop and count. Any quantity computed at a single scale is therefore a statement about that scale as much as about the world. Summing up: a measure of a nothing depends on the scale, or resolution, at which it is observed. This leads us to our first desideratum:

Scale. *A measure of nothing must be a function of the entire family $(N_r)_{r>0}$, not of any single member N_r .*

Meaningful absence, or irrelevance. Another consideration is that the zero in definition (1) is ambiguous, because $f(x) = 0$ can happen for two very different reasons. Something may be missing where it could have been: a neighborhood with no grocer. Or it may be missing where it never could have been: the open ocean contains no supermarkets, and is not a food desert for it. Only the first is a meaningful quantity, and a definition of nothing must separate the two.

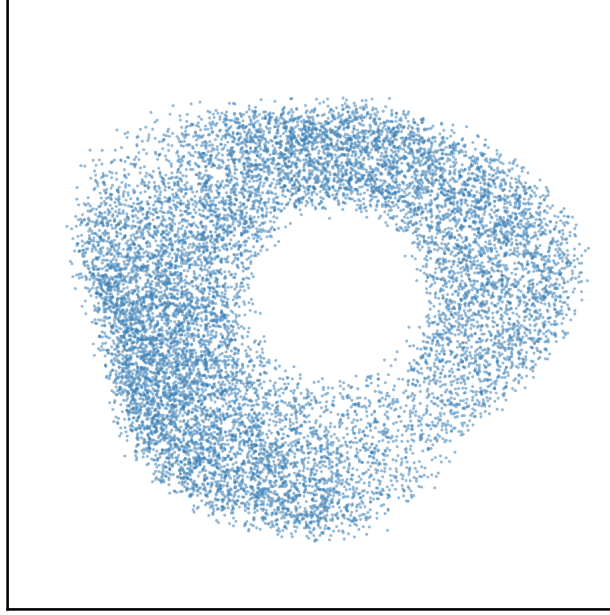


Figure 2: A point cloud with one gap: the interior pocket is enclosed by presence and is a meaningful quantity, the empty margin at the frame changes with the edge, and is not.

For a visual intuition, in Figure 2 note that there are two kinds of white spaces that we would consider ‘nothings’ without further qualifications. One is the empty circle in the center, and the other is composed of the empty edges of the figure, or the white space around the area where the data actually is. The second kind of zero, or white space, is a creature of the map, or how we define the space. Every variable simply fades off at the edge of the region where it lives: the density of grocers thins out past the city limits, gives way to fields, and is then zero for as far as one cares to look. Zeros of this kind carry no information, and they are, in some sense, arbitrary: draw the map larger and you collect more of them. The empty region at the edge of Figure 2 is like this. Its shape and size are set by where the picture was cut, not by anything in the data.

This suggests looking at what bounds an empty region. A region like the edge expanse is bounded partly by data and partly by the frame itself; redraw the frame and the region changes with it. The other white spaces in the interior of Figure 2 are different: their boundary is made of *presence*. They are closed off by the data on every side, so once they are identified, enlarging the space will not change their size. Being circled by presence settles the ambiguity of the zero: conditional on having a space in which we would expect presence, if the something exists all around an empty region, it could have existed inside it too. The empty regions that count, then, are those the data enclose. This leads to our second desideratum.

Enclosure. *Let $P_r \subset \mathbb{R}^d$ be the presence region at scale r . A connected component C of the complement $\mathbb{R}^d \setminus P_r$ is enclosed if it is bounded; equivalently, if every path from C to points arbitrarily far away must cross P_r .²*

Enclosure discards the irrelevant zeros automatically: the ocean, the fields past the city limits, the fringe of any map connect outward to the fade-off without ever crossing presence, and so are excluded, with no need to decide by hand where a supermarket “could” have been. With these two considerations in mind – scale and enclosure – we propose the following definition of a (meaningful) nothing:

Nothing. *A nothing of the data at scale r is an enclosed connected component of N_r . The nothings of the data are the enclosed components across all scales.*

Definition 2 is the estimand of this paper. Each nothing has a location, the region C itself, and, by our previous desideratum, its size must be read off the whole family: because the N_r are nested, the same pocket of absence appears at some scale, shrinks as r grows, and eventually vanishes, so that its natural size is the range of scales it survives.

Before introducing our method, it is worth asking why the familiar toolkit does not already deliver the estimand we just described. Empirical work is not silent about absence, but each of its usual devices violates at least one of the two requirements just established. Threshold indicators, a tract is a “food desert” if no supermarket lies within one mile, a citizen is “unrepresented” if no party sits within some distance, commit to a single scale r and read off N_r there; so the answer inherits the cutoff. Spatial cluster detection comes closer: scan statistics and local spatial indicators locate regions of unusually low value, but they also do so at a window or bandwidth fixed in advance, and with no notion of enclosure, so a cold spot leaking off the border and a pocket sealed by presence count alike. A more refined strategy scores access continuously rather than dichotomously: the two-step floating catchment area family (Luo and Wang, 2003), a measure in the health and food-access literatures, rates every location by the supply it can reach relative to the demand competing for that supply. But the catchment is still a radius fixed in advance (or a distance-decay curve chosen in advance, in its refinements), and the output is still a number attached to each unit: to find the deserts one must still threshold the score, and nothing in the score says whether the low-access units form a pocket sealed by presence or a fringe running off

²Section 3 and the Cambodian application return to this boundary issue and how a reference rim can close it.

the edge of the map. What we suggest below, we argue, instead follows our definition and satisfies our desiderata.

3 Topology for Data Analysis

Data as shapes. Our proposed method to quantify nothing relies on thinking about the input data differently than how empirical social scientists typically do. Taking a concrete scenario, to study food deserts for a population at risk, we might have a list of individuals and, for example, the number of supermarkets within some distance of their residence. The data would be numbers, and with these numbers, we would want to summarize them with means, correlations, measures of spread, etc. What we suggest here is instead to think of data as *shapes*. Instead of a list of counts, we would think of a food desert as an ‘empty circle’ (or other shape) in the locations of grocery stores.

For a visual intuition of what this means, consider Figure 3, which is constructed following Matejka and Fitzmaurice (2017), in turn in the spirit of Anscombe (1973). Normally, we would consider x and y as a list of numbers, and could compute their mean, standard deviation, and correlation. All six panels in the figure have basically identical summary statistics. What these would hide is the fact that the first is a blob, the second two rings, the third a circle, then a grid, diagonal lines, and an X. Our premise is that for many applications in social science, of which we will exemplify three, that shape *itself* is of interest, yet thus far it has been mostly ignored.

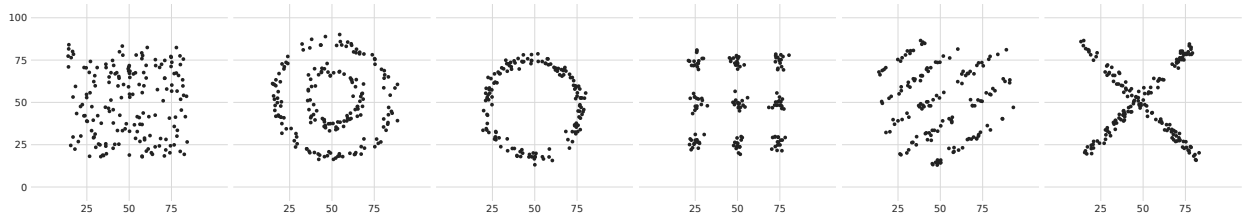


Figure 3: Six point clouds with identical summary statistics (equal mean, standard deviation, and Pearson correlation in both coordinates) but entirely different shapes. After Matejka and Fitzmaurice (2017).

Once we think of data as shapes, how can we study them analytically and quantitatively? We rely on methods developed within topology, the branch of mathematics that studies shapes, specifically of shapes irrespective of exact distances and coordinates. What we can learn are the properties of forms that are preserved even after continuous deformations (e.g. with stretching, but not tearing or gluing). In recent decades, topology applied to data analysis (TDA) has emerged as its own field, with a number of applications especially within

the natural sciences. In what follows, we will provide an intuitive overview of persistent homology. The overview is informal by design, and does not cover all TDA methods. Readers interested in more precise and technical treatments should refer to existing excellent work, which sometimes reaches the edge of understandability for non-specialists. See for instance the original Carlsson (2009), or Edelsbrunner and Harer (2010), the introduction for data analysts by Chazal and Michel (2021), and overviews in Ghrist (2008) and Fugacci et al. (2016).

Defining holes. Our measurement of nothings will rely on measuring holes; and topology measures shapes by counting their holes. As it is intuitive, and as the previous section foreshadowed, the definition of a gap or a hole cannot be detached from the number of dimensions the data lives in, see Figure 4. The formal tool for this counting is *homology*: to each shape it assigns, in every dimension k , an integer β_k , the k -th Betti number, equal to the number of independent k -dimensional holes it contains. A zero-dimensional hole is the gap that separates one piece from another, so the count of zero-dimensional features is the number of connected components. A one-dimensional hole is a loop that encloses an empty region, such as the hole in a ring. A two-dimensional hole is a void enclosed by a surface, like the cavity inside a ball. The idea continues into higher dimensions, where the visual intuition stops but the arithmetic does not, so that a shape of any dimension has a well-defined count of holes at each level. While technically possible in any dimension, in all of our practical applications we would consider one-dimensional holes, meaning a populated or dense region in two-dimensional space surrounding an empty one.

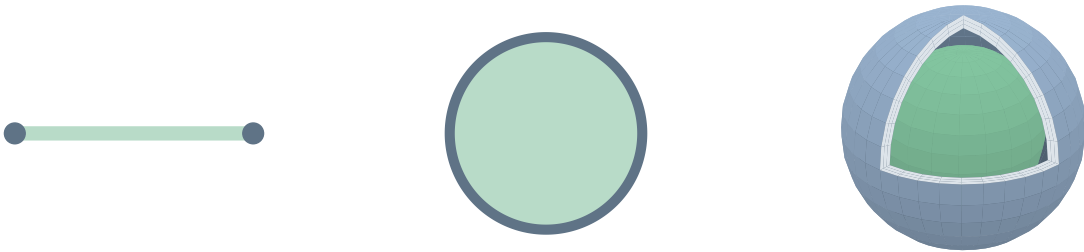


Figure 4: A gap, in dimensions zero, one, and two. In each panel the data (slate) surrounds a region it does not fill (green): two points enclose the segment between them (dimension 0, $\beta_0 = 2$ components, and so one gap, since n components leave $n - 1$ gaps between them); a circle encloses a disk (dimension 1, $\beta_1 = 1$ loop); and a sphere, drawn with one octant cut away to show its interior, encloses a ball (dimension 2, $\beta_2 = 1$ void).

Building a shape from data. Starting with two-dimensional data, we would have a scatter of points in space: that has intrinsically no holes, it would just be separate dots. How does one measure the shape of the data, then? To do so, we have to ‘connect the dots’ into a continuous object, and the standard way is to build a *simplicial complex*, i.e., a figure assembled from vertices, edges, filled triangles, and their higher analogues, as illustrated in Figure 5. Once the data is expressed as such a complex, the holes in every dimension can be computed exactly; the real choice is how to build such a complex, or, in simpler terms, which edges and triangles to include.

More precisely, the pieces of a simplicial complex are called *simplices*, one for each dimension: a single point is a 0-simplex (a vertex), a segment joining two points a 1-simplex (an edge), a filled triangle on three points a 2-simplex, a solid tetrahedron a 3-simplex, and so on into higher dimensions. These pieces are connected along shared faces, and the collection is required to be closed, meaning that whenever a triangle is included, so are its three edges and three vertices. The top row of Figure 5 shows the lowest of these building blocks, a vertex, an edge, and a filled triangle, and the panel below shows a complex assembled from them: a set of vertices joined by edges into filled triangles, which here surround an empty region in the middle, a one-dimensional hole. The advantage of the construction is that, on an object built from such simple pieces, the holes of each dimension can be counted by elementary linear algebra, so that once the data has been turned into a complex, its shape becomes something we can actually compute.

Filtrations. This then opens the question of *how* to connect the dots. As discussed in Section 2, scale matters for defining a gap, and the same issue is a key part of the intuition here as well. Suppose you had a point cloud in two dimensions: you could fix a radius r , draw a ball of that radius around each point, and join points whose balls overlap. You could fill in an edge for each overlapping pair, a triangle for each mutually overlapping triple, and so on. That would give you a way to connect the dots, but what would be the correct r ? Too small a radius leaves the data as disconnected dots; too large a radius merges everything into one featureless blob; and the features the researcher wants to capture, and are genuinely present, may appear only at intermediate scales. Rather than picking one radius, the intuition here is to consider all of them at once: as r grows from zero, the complex only ever gains pieces, giving a nested, growing sequence of complexes, each contained in the next. This sequence is called a *filtration*.

The particular way of connecting the dots is itself a choice. If a simplex is added as soon as its points are pairwise within reach, it is called the *Vietoris–Rips* filtration: two points

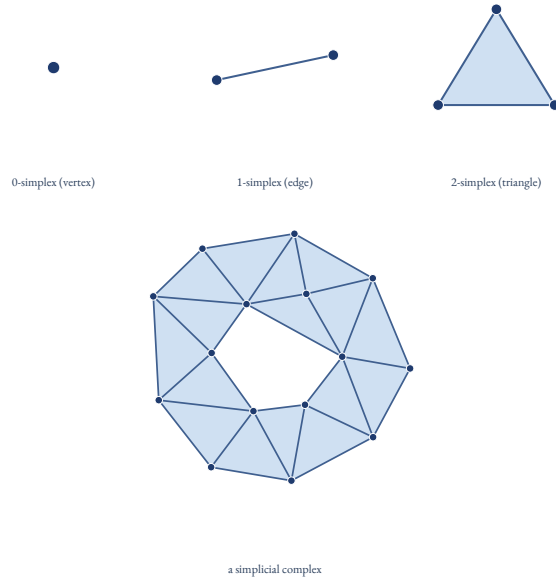


Figure 5: The building blocks of a simplicial complex. A vertex (0-simplex), an edge (1-simplex), and a filled triangle (2-simplex), and, below, a complex assembled from them: vertices joined by edges into filled triangles, connected along shared faces, here leaving a one-dimensional hole in the middle.

are joined by an edge when they are within reach of one another, and a triangle is filled as soon as its three sides are present, with no further condition checked. Because the rule only ever asks whether pairs are close, the whole construction is determined by the list of pairwise distances, needing no coordinates and no geometry, which makes it simple and cheap to build, and the standard tool when the data are a bare point cloud whose only structure is the distances between points. A close relative is the *union-of-balls*, or *Čech*, filtration, pictured in Figure 6: here we literally grow a ball of radius r around each point, the shape at scale r is the union of those balls, and a group of points is filled in only when its balls overlap *jointly*, not merely two at a time; a hole closes exactly when the balls around it have grown enough to cover it, so that as r increases the union fills in and its holes disappear one by one. This version follows the geometry faithfully, and is the natural choice when one wants the filtration to track the union of balls itself.³

The examples we use below, which we think are representative of the type of challenges social scientists face, however, do not start with a bare point cloud. The data comes with a notion of neighborhood (specifically, below: census blocks that share a border, districts adjacent on a map, and cells of a lattice over ideological space). They also come with a

³In practice it is computed efficiently through the closely related *alpha complex*.

measured quantity attached to each unit, such as the density of a population, a sex ratio, or the distance to the nearest party. In that setting we build the complex directly from the given adjacency, and instead of growing a geometric radius we sweep a threshold through the measured field, a *sublevel-set* (or, sweeping the other way, *superlevel-set*) filtration: we admit units in order of that value, and watch the shape fill in as the threshold rises (or falls). The intuition is the same as growing balls, only the parameter being swept is a substantively meaningful variable rather than a pure distance, and it is the choice of that quantity that fixes what an ‘empty’ region will mean. In other words, the role played by the growing radius in the union of balls is played here by the threshold: a block is added to the shape once the value of its variable (e.g. population density) reaches it, and we consider all thresholds.

Persistent Homology. We have now a way to organize data as shapes through filtrations, and count gaps, or the existence of nothings, through homology. The goal of our method is not only to detect the existence of gaps, but to quantify their sizes. We will combine the two ideas to do so. Homology, as above, counts the holes of a single complex; *persistent* homology tracks those holes across the whole filtration. As the parameter grows and the shape fills in, a hole of a given dimension appears at some value (its *birth*), and later closes up and disappears at some larger value (its *death*). The difference between the two, the *persistence* of a hole, records how much of the sweep it survives. This is a key quantity: if a hole is born and dies immediately, it could be either noise (or the product of a given scale), or a gap at the micro-level. If a hole persists across a wide range of parameter space (distances or values), that would suggest either a proper signal, or a feature happening at a more macro-level. Tracking the persistence of holes is generally displayed as a barcode, one bar per hole running from its birth to its death, or as a *persistence diagram*, one point per hole placed at its birth and death coordinates. In both representations the long bars, or the points lying far from the diagonal, represent the larger features. It is this persistence, the size of a hole measured by how long it survives the sweep, that we will use to quantify our nothings.

To make this concrete, Figure 6 walks through the whole process on a point cloud sampled from three loops, one large and two small. Panels (a) to (f) show the union of balls as the radius grows, each next to the barcode assembled up to that point. At the start (a) there are only isolated points, and the barcode is a set of short bars, one for each component, tracking dimension zero (connected components). As the balls grow (b), nearby points connect and components merge, so those bars end one after another. The features we care about appear next, in dimension one (loops): the hole of the large loop closes first (c) and is joined a little later by the two small loops (d), each new hole opening a new bar. As the balls keep growing,

the two small holes fill in and die (e), while the large one, being wider, survives much longer before it too closes and the shape becomes a single blob (f). The large loop is therefore the most persistent, born early and dying late, and it leaves the longest bar; the two small loops leave short bars, and the merging components leave the many short bars that end near the start. Panel (g) places the completed barcode beside the persistence diagram. Notice that the two carry the same information: every hole appears once as a bar, running from its birth to its death, and once as a point, placed at its birth and death coordinates. In both, the persistent hole is the one that stands out, as the long bars or as the points lying far from the diagonal.

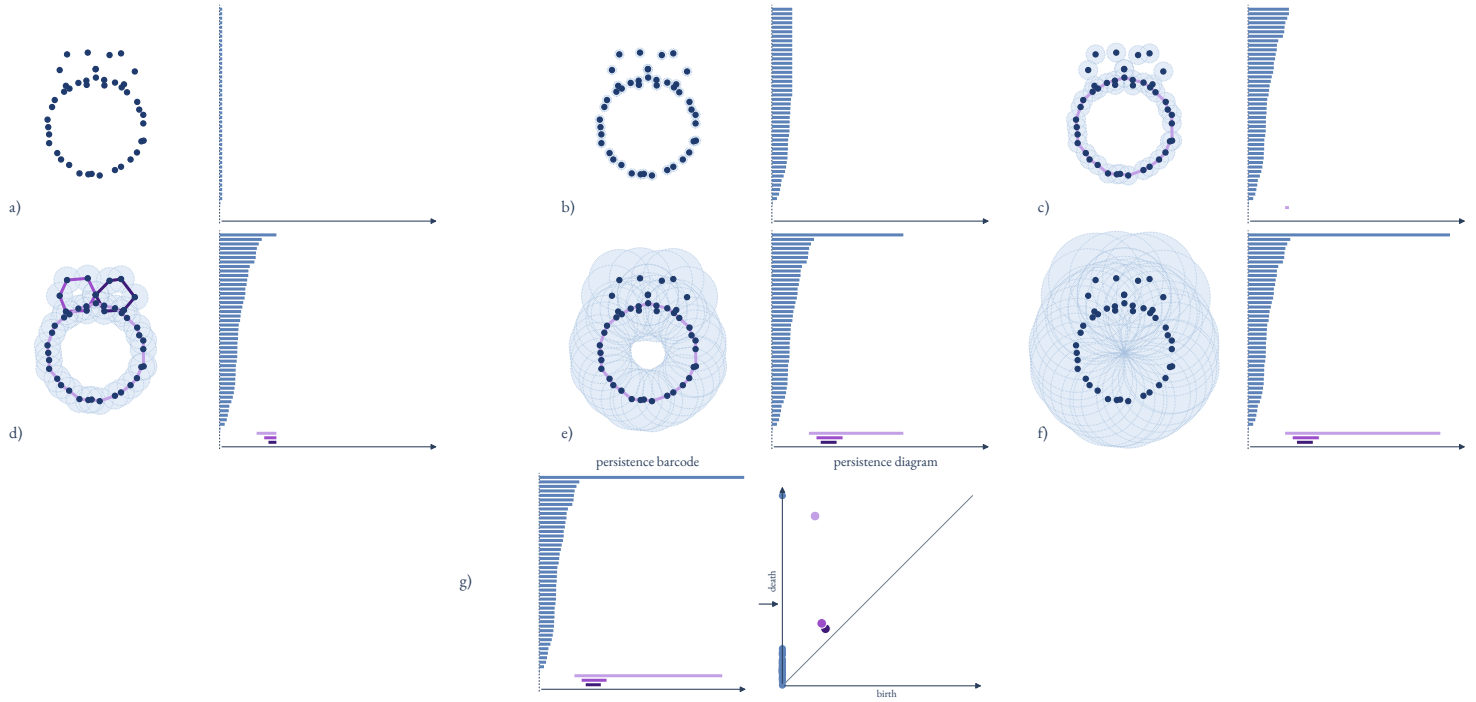


Figure 6: Persistent homology on a point cloud, adapted from Chazal and Michel (2021). Panels (a) to (f): as the radius of a ball around each point grows, the union of balls (light blue) fills in, and beside each stage is the barcode built so far. Panel (g): the completed barcode and the persistence diagram, where each hole appears as a bar from its birth to its death, and as a point at its birth and death coordinates. The persistent holes are the long bars, equivalently the points far from the diagonal.

More on implementation. For practical purposes, it is worth introducing two more concepts, that we use in our applications below and that are potentially useful to other researchers. First, one should note that the diagram just described is a set of points of no fixed size, which is awkward to compare across cases, or to feed into ordinary statistics. A common remedy converts each diagram into a *persistence landscape*, a set of curves built from

the diagram that lives in a space where averaging and distances behave normally, so that diagrams can be summarized by a single number or compared across groups while keeping their meaning.

Second, some absences are defined by two conditions at once. For instance, a food desert is not just the absence of grocery stores: we would not consider a real desert a food desert. It is instead the presence of population *and* the absence of grocery store. A sweep over a single variable cannot capture such a combination, so one sweeps two thresholds together and looks for holes that persist across a range of *both*. The intuition is unchanged, only the computational cost is heavier. This process of sweeping two parameters at once is called a *bifiltration*. The two-parameter case has no clean decomposition into birth-and-death intervals, and hence no equivalent of the barcode shown above (Carlsson and Zomorodian, 2009). The analysis relies on other summaries, such as a count of how many holes are present at each pair of thresholds (the Hilbert function) or an average of a hole’s presence over the whole range of thresholds. Botnan and Lesnick (2023) survey this area.

One practical caution runs throughout: the computation is cheap for components and loops but grows expensive fast for cavities in three dimensions and higher-dimensional holes, whose number explodes (Otter et al., 2017).

Stability and Inference. The reading rule above, that persistent holes are real and short-lived ones are noise, comes from a firmer basis than eyeballing distance from the diagonal. The foundation is a stability result: persistence diagrams change only a little when the data change a little, so that a small amount of measurement error moves each point only slightly and cannot conjure or erase a persistent hole (Cohen-Steiner, Edelsbrunner and Harer, 2007). Stability is what allows the claim that high-persistence features reflect the underlying population rather than the particular sample. Note, however, that this is a statement about small, uniform perturbations. It is not a statement about outliers not mattering. Adding an outlier (meaning a point that is far from the rest) would naturally change the features of the shape of the data.

Another natural question relates to how to do inference with the features extracted through persistent homology. The standard approach is resampling: by repeatedly subsampling the data and recomputing the diagram, one builds a confidence interval around it, a distance from the diagonal below which points are indistinguishable from noise at a chosen level, so that only holes reaching above the band are declared significant (Fasy et al., 2014; Wasserman, 2018). Comparisons between groups, or against a null model, can be handled by permutation along the same lines. Chazal and Michel (2021) discuss existing bootstrapping procedures, and we plan to show implementations thereof in future iterations of the present

draft.

Summary. To sum up the whole procedure, the study of gaps through persistent homology occurs as follows. We build a simplicial complex on the space the data occupy; we choose what to sweep, a distance, or one or two (or more) substantive variables at once; we grow the shape, record which holes are born and how long they persist, and interpret the more persistent ones as the larger gaps. It is worth pausing to note that this machinery delivers the desiderata of Section 2. Homology registers a gap only when it is enclosed, so absence surrounded by presence is detected from the data alone, with no pre-specified list of candidates, and each hole is an individual object that can be placed back in the original space. Persistence sizes every hole on the same ruler, and because the filtration sweeps all thresholds rather than fixing one, the resulting measure is scale-free: gaps can be ranked within and across cases without committing to a radius, cutoff, or areal unit.

Before turning to the three substantive applications, we illustrate the entire procedure on a simple, well-known space with a single variable, where it is well known what the largest nothing is.

Manhattan as a proof of concept. In the following, we aim to show how persistent homology identifies Central Park as the largest residential gap in Manhattan. Indeed, if it measures what we claim, it should find an eight-hundred-acre rectangle with almost no residents, rank it as the deepest gap in the island, and place it correctly on the map, with *no guidance* beyond the population counts. Figure 7 shows that it does, and in the process walks through every step of the previous section.

We start with a field of the population density of Manhattan for each census block.⁴ The top-left panel of Figure 7 displays that field, bright where people are packed most densely and dark where they are not. Onto this field we run a superlevel sweep, defined above in the discussion of filtrations. We start from a high density threshold, so that only the most crowded cells count as “populated,” and lower it step by step, admitting sparser and sparser cells until the whole island is included. A void is a low-density pocket enclosed by a denser rim.⁵ Its persistence, birth minus death, measures how deep the population dip is relative to the neighborhood that encircles it.

⁴We begin with the 2020 census block populations for Manhattan, which sum to the borough’s official count of 1,694,251 residents. We project the blocks to a metric coordinate system, lay a one-hundred-meter grid over the island, add each block’s population into the cell that contains it, and smooth the result into a continuous density field measured in persons per square kilometer.

⁵To further clarify, a void is born at the moment the threshold falls below its surrounding rim, when the ring of dense cells around it closes into a loop, and it dies when the threshold falls below the pocket’s own interior floor and the hole fills in. A void’s birth density is therefore the density of its rim and its death density the density of its floor.

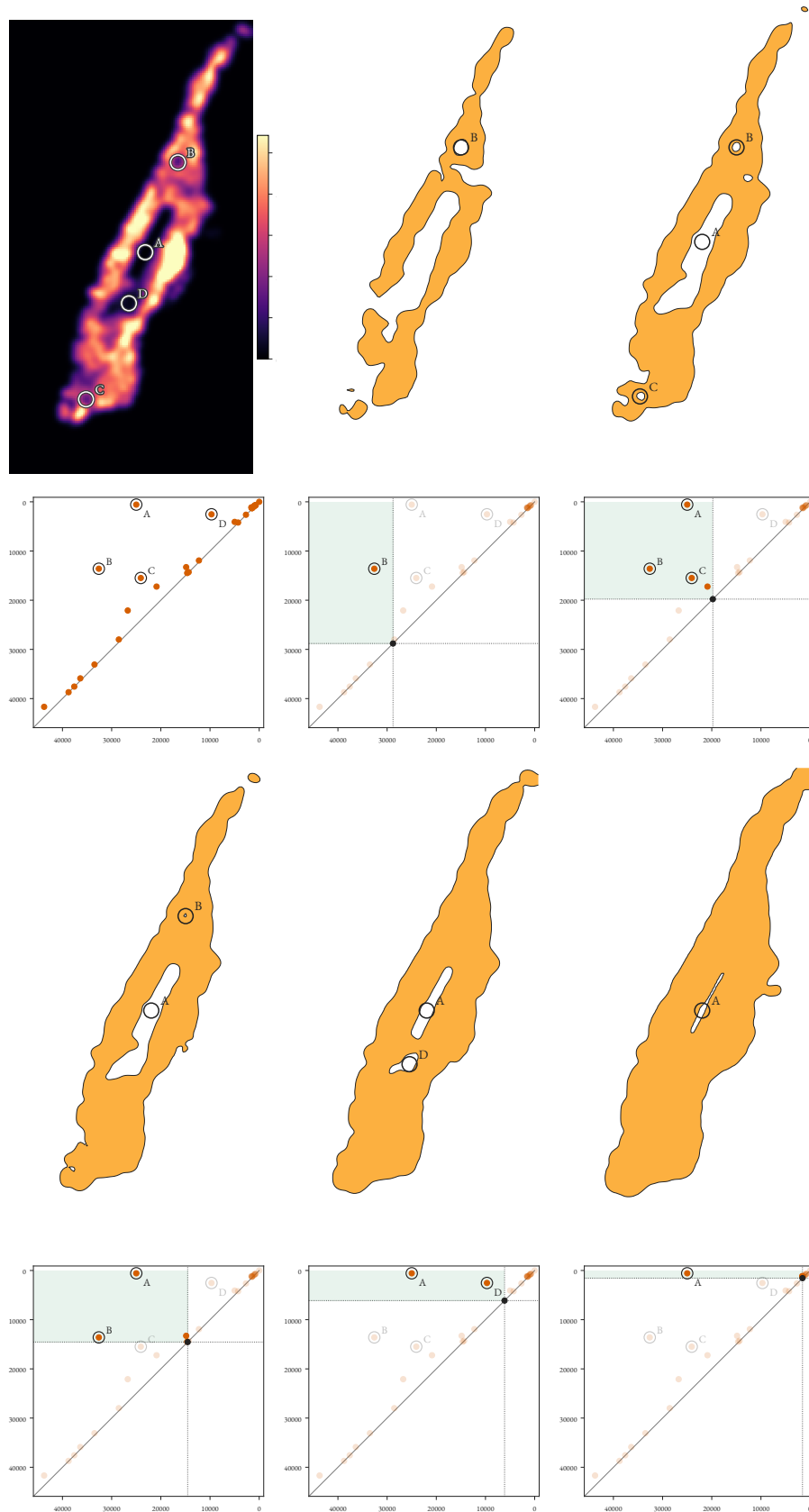


Figure 7: Manhattan's residential-population voids. Top-left: the reference population density and the persistence diagram of the smoothed block-population field, the four most persistent voids circled. Remaining figures: the populated region as the density threshold falls from 28.8k to 1.6k persons per km², and resulting appearing and disappearing gaps.

The persistence diagram in the upper-left panel records one point per void, with four points standing well clear of the diagonal, indicating more persistent (larger) holes. Localizing those four voids places each on the map. In order of persistence, they are: Central Park (A), the seam between the City College campus and St. Nicholas Park (B), the civic center around City Hall Park (C), and the Midtown office core (D). Central Park is in a class of its own: a near-total absence of residents ringed by some of the densest housing in the country, and hence by a wide margin the most persistent hole in the field. The remaining columns of Figure 7 make the lifecycle concrete by showing the populated region at five descending thresholds.

Note once again that the method was given nothing but block populations. It was never told that Central Park exists, let alone where, yet it recovered the park as the deepest gap. This is what it means to measure an absence rather than to confirm one we had already marked. With the procedure seen to work where we can check it by eye, we turn to three illustrative settings in which the gaps are known to exist or visually identifiable, but less trivial to measure.

4 Applications

We next illustrate the method with three applications. Our aim here is not to advance new empirical findings, but to show that and how the method works. The cases were therefore chosen such that their answers were already known in broad outline, or open to direct visual inspection. At this stage, what is new are not necessarily the conclusions, but the measurement itself. That the exercise is possible at all is itself the point: in each case we measure the size and location of an absence that, to the best of our knowledge, no other method could quantify systematically.

4.1 Food Deserts

Background. The most natural application of our method to measure nothings, holes, and gaps is the lack of access to resources for certain populations. Within this broader category, access to nutritious, fresh and affordable food and the existence of ‘food deserts’ is notoriously a problem in the US, and possibly elsewhere. Yet despite three decades of active research, the empirical literature on food deserts is still divided. Findings reverse across studies, and this is in large part due to methodological instability. There is no agreed definition of access, studies vary in the store types they count, the buffer distance that separates “served” from

“underserved,” and the spatial unit of analysis (Lytle, 2009; McKinnon et al., 2009), and systematic reviews report that the evidence for food deserts is mixed across countries and sensitive to these choices (Beaulac, Kristjansson and Cummins, 2009; Walker, Keane and Burke, 2010; Apparicio, Cloutier and Shearmur, 2007).⁶

In sum, much of the empirical instability is a version of the modifiable areal unit problem: results change with the scale and the units at which access is measured. Framing the desert as a “nothing” to be measured topologically addresses this issue. Instead of fixing a pre-specified radius, and asking whether a given tract is a desert, we measure the gap itself, at every scale at once, so that the answer does not depend on an arbitrary cutoff.⁷ We hence apply persistent homology to the presence of grocers relative to the population that depends on them, and we focus on selected US cities as an illustrative case study.

Data. Since the lack of access to fresh food is especially salient for the poorer strata of the population, we focus on participants of the Supplemental Nutrition Assistance Program (SNAP). For demand, this is the population most constrained by local food access, less able to absorb travel costs, and less able to substitute across stores. For supply, SNAP-authorized general-line stores are those that actually accept their benefits. The supply layer comes from the USDA SNAP Historical Retailer Locator, restricted to grocery types that are not convenience stores (supermarkets, super stores, and large, medium, and small grocery stores) active in 2020.⁸ The population layer is the number of SNAP-receiving households per square kilometer (ACS 2016–2020), attached to 2020 census blocks.⁹

Implementation. We treat each metropolitan area as a field with census blocks, where every block has a distance to the nearest grocer, and a SNAP-household density. We then build a graph of the centroids of the blocks,¹⁰ and on this graph we run persistent homology. Remember from Section 3 that what this gives us is the existence of holes, and their persistence (or ‘size’). One could be tempted to simply look at the relative position of SNAP grocers

⁶Note that regardless of whether food deserts are driven by supply or demand (Allcott et al., 2019), the measurement problem still stands.

⁷This does not imply, however, that we are not constrained by the level or unit at which the data is collected.

⁸We deliberately exclude the “fringe food” tier, such as convenience stores, drugstores, dollar stores, and combination grocers, that dominates the food-stamp retail counts but does not sell a full grocery basket, to focus on fresh foods.

⁹SNAP-receiving households are reported by census block group, which we assign to census blocks proportional to number of households.

¹⁰Specifically, we connect the blocks into a graph by a Delaunay triangulation of their centroids, and we verify that they form a single connected component. This is a necessary safeguard, since an isolated block would otherwise be mistaken for a fully enclosed gap, which would not be of interest.

across census blocks, and identify gaps among them. This, however, would not produce the quantity of interest: gaps could simply be due to geography, such as to rivers or parks, or coincide with rich neighborhoods, that tend to have no SNAP grocers or few grocery stores in general. What a food desert should capture instead is an area with no SNAP grocers *and* high density of SNAP recipients.

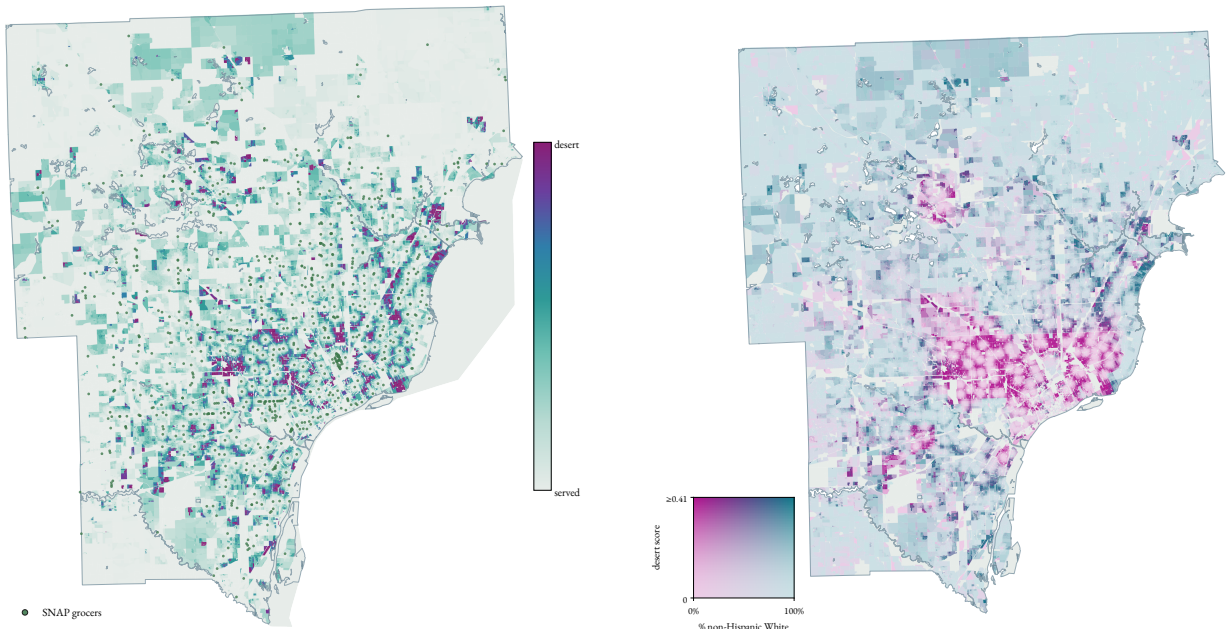
Our measure of gaps is hence defined both by distance (how far one is from a grocer) and density (how much we should expect there to be a grocer). Using a bifiltration, a block is then acceptable, or served, when it is close enough or sparse enough; it is *underserved* when it is instead far and densely SNAP-populated. The topological features of interest are the one-dimensional holes (H_1) in this served-underserved continuous measure of ‘space’. A bounded cluster of underserved blocks, each both far from a grocer and densely populated by SNAP households, completely encircled by served blocks, is a loop enclosing a gap. This is, literally, a food desert: an underserved island ringed by served or empty land. Further following Section 3, the persistence of such a hole (the range over which it survives before the served region closes it off) measures its severity. Going back to the geography challenge mentioned above, consider a remote but empty pocket, such as a park or an airport: with our approach, it does not introduce noise, as it dissolves as soon as the density gets to near-zero population, so it registers as insignificant automatically, with no hand-tuned cutoff. Crucially, we do not choose a single distance threshold, density threshold, or exchange rate between the distance and density. Instead we integrate each block’s enclosed-desert membership over the entire range of values of the two, yielding a per-block “desert score.” Sweeping over all these values is what makes the measure scale-free: it reports the gap structure across all resolutions at once rather than at one chosen scale.¹¹

Validation. We report the results for a few key cities to illustrate the output.¹² Figures 8 and 9 report their maps with SNAP grocers (green dots) and food deserts overlaid. Blocks are shaded from the basemap color (“served”) toward magenta in proportion to their desert score. Figure 8 shows the Detroit tri-county lattice, both in the simple form of SNAP population without considering its race (Figure 8a), as well as with hues encoding a block’s non-Hispanic-White share (Figure 8b), so that the racial geography of the deserts is legible alongside their severity. Figure 9 places Phoenix and Philadelphia in identically-sized windows; because these two panels are cropped from a single national sweep and share one color ceiling, their intensities are directly comparable.

¹¹One technical choice matters: the two axes are rank-normalized to empirical percentiles before the sweep, so that the heavy-tailed density axis cannot silently dominate the far-and-dense conjunction.

¹²We have run the same analysis for the main metropolitan areas of the US, which we are happy to share under request, and will eventually be put in an Online Appendix.

We chose Detroit as an example to validate our method, since it is one of the best studied cities for food deserts. A preliminary analysis of the data suggests that our results are in line with Zenk et al. (2005). The enclosed-desert mass concentrates in the city and its inner ring, and a racial component is clearly visible, with food deserts tending to correspond to blocks with a low share of white residents. When our scores are run through Zenk’s own design, based on poverty-by-race tertiles, they reproduce the poorest-tertile racial disparity (a contrast of roughly one standard deviation, $p < .001$). The result is thus consistent with the canonical finding, supports the location of the food deserts, but it improves on it in an important way: we never stick to Zenk’s fixed distance, or to the tract as the unit of analysis.



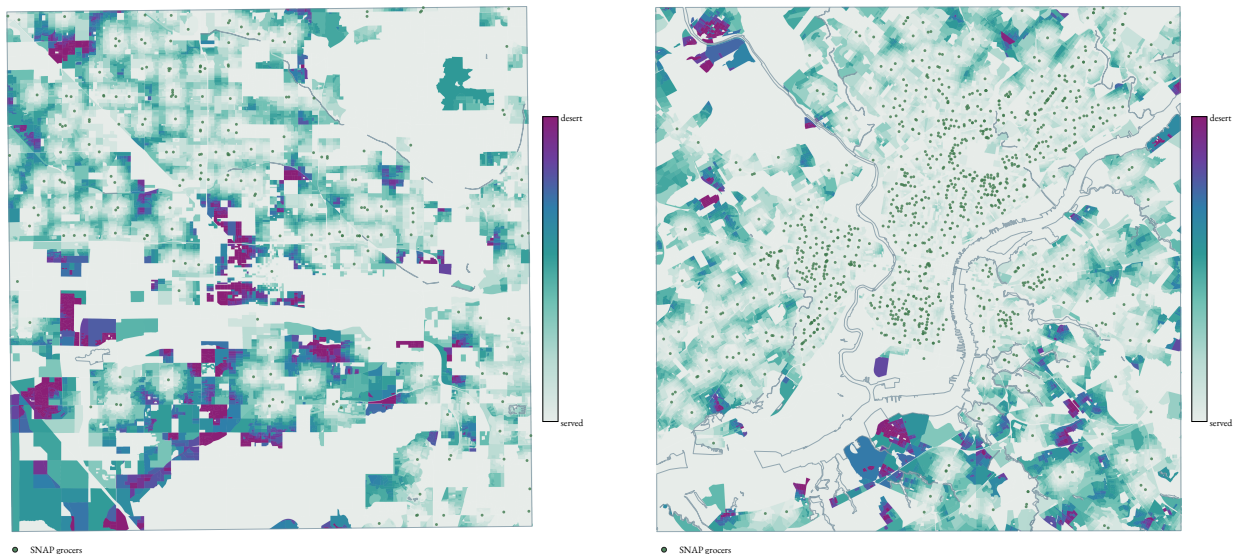
(a) Detroit Food Deserts, SNAP Population.

(b) Detroit Food Deserts, SNAP Population by Race.

Figure 8: Food deserts in Detroit, on the tri-county lattice (Wayne, Oakland, Macomb). Green dots are SNAP-authorized grocers; each census block is shaded from the basemap color (“served”) toward magenta in proportion to its desert score, the enclosed-desert membership integrated over all distance and density thresholds. Panel (a) shades by desert score alone; panel (b) overlays each block’s non-Hispanic-White share.

As an additional consideration, comparing Phoenix and Philadelphia in Figure 9 makes the imprint of urban form clear, especially the differences in urban development between the East and West of the United States. Philadelphia is compact: a dense, pre-war fabric in which grocers are numerous and close, so that travel times are short and food can be reached by a brief trip; access gaps are correspondingly not severe. Phoenix, on the other hand, has

a postwar, automobile-oriented, gridded development, clearly visible in the figure, which spreads a sparser set of large-format grocers across long distances, so that reaching food effectively requires a car.¹³ The food-accessibility consequences of path-dependent urban development thus manifest in our measurement, without resorting to an arbitrary threshold or cutoff.



(a) Phoenix Food Deserts, SNAP Population.

(b) Philadelphia Food Deserts, SNAP Population.

Figure 9: Food deserts in Phoenix and Philadelphia, with SNAP grocers (green dots) and per-block desert scores (magenta).

4.2 Genocide

Background. Between April 1975 and January 1979, the Khmer Rouge governed Cambodia as Democratic Kampuchea. Pursuing an immediate and total agrarian communism, the regime emptied the cities, abolished money, markets, and religion, and drove the population into rural labor collectives. The death toll from execution, starvation, overwork, and disease is commonly estimated at between roughly 1.5 and 2 million people, on the order of a fifth to a quarter of the population (Heuveline, 1998; Kiernan, 1996), making it one of the most brutal genocides in recent history. The killings were not indiscriminate. The regime targeted the urban and educated, former officials and soldiers of the ousted republic, ethnic and religious minorities, such as Cham Muslims, ethnic Vietnamese and Chinese, and Buddhist monks.

¹³The wide zero-score band through the center of the Phoenix window is not a served neighborhood but has no population, the airport, the Salt River bed, and Papago Park. So it is correctly not identified as a desert.

Men were disproportionately targeted, as they were more likely to be executed as soldiers, officials, or suspected enemies (Kiernan, 1996).

A substantial demographic literature has reconstructed the resulting loss, and in particular the deficit of men. de Walque (2005, 2006) documents selective mortality by sex, age, and education; others trace the skewed sex ratios and the “demographic scar” that the cohorts alive under the regime still carry decades later (Heuveline and Poch, 2007; Neupert and Prum, 2005; Sliwinski, 1995; Banister and Johnson, 1993), and recent work links the imbalance to long-run political outcomes (Gaikwad, Lin and Zucker, 2023). We revisit this record for two reasons. First, our method offers a way to measure these gaps directly, by locating, in space, where men are missing, and to size the deficit relative to its surroundings. Second, and precisely because the deficits are already known to exist and are corroborated by an independent record of documented mass graves, the exercise is also meant as validation. The goal, in other words, is to detect these “nothings” in the right cohorts and the right places.

Data and Implementation. We use IPUMS International census microdata for Cambodia for 1998, 2008, and 2019. From each census wave we form person counts, for every district $\times$ sex $\times$ five-year birth cohort, placing each person by district of birth.¹⁴ Each district then has for each cohort a smoothed sex ratio $(M + \frac{1}{2})/(F + \frac{1}{2})$. Then, we use the districts as the nodes of a spatial complex, and we join neighboring districts by edges and fill mutually adjacent triples as triangles. We augment them with a rim of foreign districts that supplies an ‘un-genocided reference’ against which a deficit at the national border can still register.¹⁵

With the complex in hand, we can run persistent homology, and we do so separately for each cohort and census. Similarly as above, we again consider one-dimensional holes (H_1). In this application, a hole or a desert is a district where men are missing relative to women. The hole persists until the sex ratio would be restored, so that persistence in this case captures the severity of the imbalance of the sex ratio, in a given district relative to its surroundings.

The persistence diagram. Per Section 3, the raw output of persistent homology is a persistence diagram for each cohort, which we stack and plot in Figure 10 for the 2008 census.

¹⁴We decided to place people based on birth place rather than residence, as the post-1979 migration and urbanization would introduce noise relative to the population distribution at the time of the Khmer Rouge.

¹⁵The rim (“collar”) is made of adjacent foreign provinces in Laos, Thailand, and Vietnam whose counts are age-shifted to the Cambodian census year. Where the national boundary meets the sea and has no foreign neighbors, the coast is closed by triangulating each coastal district onto its nearest collar unit and cross-linking adjacent coastal districts, so that a border pocket is enclosed rather than leaking off the edge of the map.

In this specific case, a point is one of the holes or gaps just described. The distance from the diagonal indicates its persistence. We indicate younger cohorts with lighter colors. The interpretation of the diagram is straightforward: if points are close to the diagonal, they indicate shallow, or ‘short-lived’ holes: they appear and disappear quickly, meaning the sex ratio is close to 50:50. The figure shows that this is clearly the case for the younger cohorts. On the other hand, if points are further from the diagonal (closer to the bottom), they indicate deeper holes, or a stronger deficiency of males. The genocide, in other words, is legible in the shape formed by these points, where the high-persistence points are concentrated in the older cohorts who would have been alive during the Khmer Rouge years.

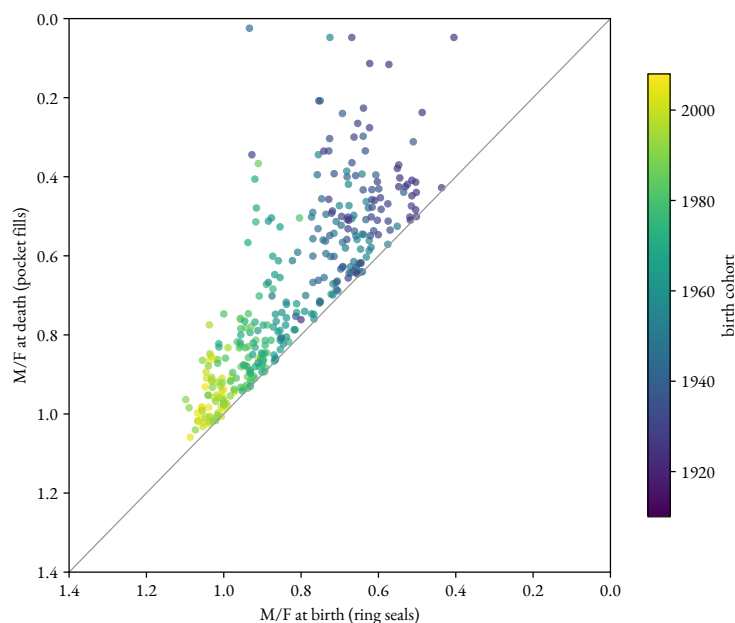


Figure 10: Persistence diagram of missing men in the 2008 census. Each point is a district-level hole in the male-to-female sex ratio, color runs from lighter (younger cohorts) to darker (older). The distance from the diagonal measures persistence, i.e., the severity of the male deficit relative to surrounding districts.¹⁶

Demographic gaps by cohort. For a clear view of the full demographic distribution, and which cohorts have more missing men, Figure 11b shows the depth of holes in the sex ratios across census waves. The same figure is done for neighboring Laos in Figure 11a, for reference. The comparison shows how anomalous Cambodia’s demographic gap is. To construct a single number for each cohort for the figures just described, we transform every cohort’s persistence diagram into a persistence landscape. This means stacking the holes into a set of curves, whose height and spread encode how many deep features the cohort contains. We then take the L^2 norm of that profile to turn it into a scalar (a single summary number).

Figure 11 arranges these norms as a heatmap: rows are five-year birth cohorts on a fixed grid, columns are the successive censuses, and color is the norm. Darker colors indicate that bigger holes (missing men) are present in that cohort. Cambodia (11b) shows gaps across the 1925–65 cohorts that persist across the 1998, 2008, and 2019 censuses: a demographic scar that ages with the people who carry it. Laos, too, produces high norms, but they are found in each census’s oldest observed cohorts and move as the census moves. That being said, Laos is still an imperfect reference, as we can only locate its people by residence, and it had its own conflict in the 1960s and 70s. But the contrast is striking: only Cambodia carries a stationary cohort gap, in the cohorts the genocide predicts. Next, we check whether these gaps are also in the *places* the genocide predicts.

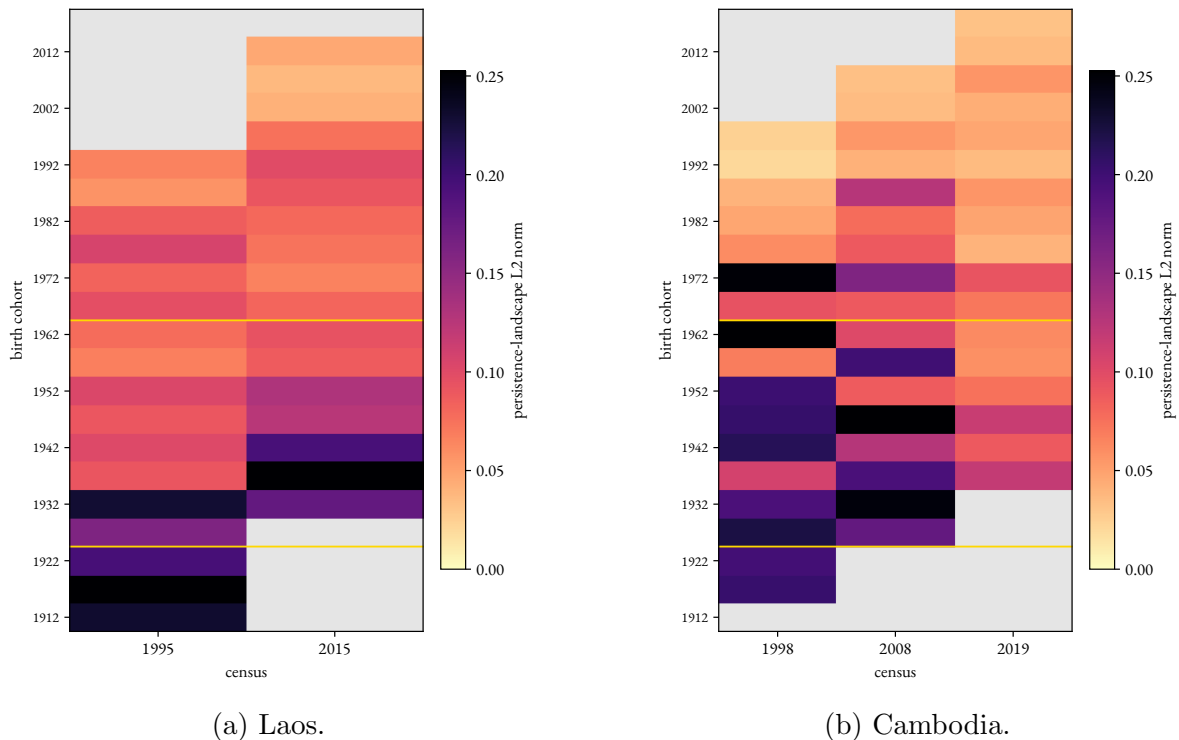


Figure 11: Persistence landscapes of the sex-ratio holes. Rows are five-year birth cohorts on a fixed grid, columns are the 1998, 2008, and 2019 censuses, and color is the L^2 norm of each cohort’s landscape, darker where deeper holes (missing men) are present. Cambodia (b) carries a stationary gap across the 1925–65 cohorts that holds fixed across all three censuses. Laos (a), for reference, shows high norms only in each census’s oldest observed cohorts, which move with the census rather than tracking a fixed birth cohort.

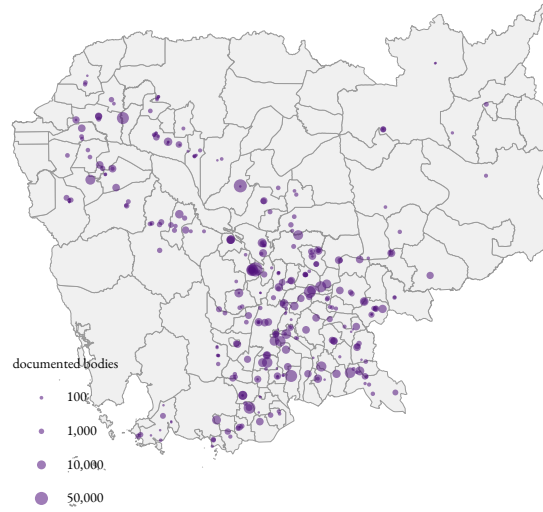
Spatial distribution of the demographic gaps. We locate where the holes are by visualizing the demographic gaps by district in Figure 12. The graves panel (12a) plots the mass-grave sites documented by the Documentation Center of Cambodia. These 309 locations are sized

on the map by the number of recorded bodies. This uses completely independent evidence to mark where the killing was heaviest. The Khmer Rouge panel (12b) pools the gap intensity of the exposed cohorts (born 1925–65) across all three censuses. These are the demographic scars our method recovers. The third figure, (12c) repeats the exercise for all other cohorts on the *same* color scale, and serves as a benchmark. This is what the country’s sex-ratio geography looks like absent the shock. The contrast shows that the exposed cohorts have gaps that are stronger and more spatially concentrated, especially where the graves are. The unexposed cohorts, on the other hand, report a fainter, more diffused pattern.

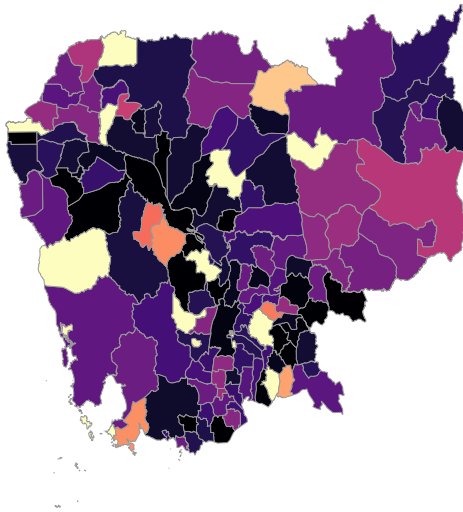
4.3 Representation

Background. Next, we apply the same method to one of the key questions in political science: what does representation look like in a democracy? Who is represented, and who is not? A long tradition treats representation as congruence, i.e., the correspondence between the positions or ideal points of citizens and those of the parties or legislators who stand for them (Pitkin, 1967; Achen, 1978). Operationally, congruence is measured through distance in ideological space: a citizen is well represented when a party is close to their ideal point, and a polity is congruent when the distribution of parties tracks the distribution of voters (Huber and Powell Jr., 1994; Golder and Stramski, 2010). Much of the empirical literature has largely summarized this correspondence as an average, such as the mean citizen–representative distance, or the gap between the median voter and the median legislator.

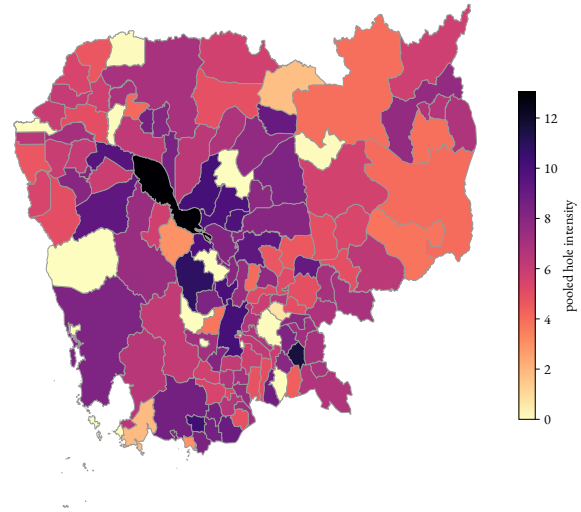
For our application, we measure individual voters and parties’ positions (as opposed to politicians within parties). In a two-party system, with only two ideal points for the parties, understanding the representation gap, or which regions in space are not occupied by parties is essentially trivial. We hence focus on the multiparty proportional systems in Europe, illustrating it with the cases of Germany and Italy, and considering a two-dimensional space with a socio-economic component and a cross-cutting component. What this means is that we could potentially establish whether an economically left culturally conservative voter (for example) is not represented by any party (Lefkofridi, Wagner and Willmann, 2014). It is there that a representation gap is substantively interesting. Yet, a standard average-distance measure cannot see it: it reports how well the typical voter would be served, not which regions of preference space are served by no one, nor how large those unserved regions are. Reframing a representation gap as a topological “nothing”, i.e., a hole in the party coverage of a populated region, lets us locate the unrepresented regions directly, size them, and rank the larger gaps against the smaller ones, across the whole space and at every scale at once,



(a) Mass Graves.



(b) Genocide Cohorts.



(c) Other Cohorts.

Figure 12: Genocide locations and demographic gaps. Panel (a) plots the 309 mass-grave sites documented by the Documentation Center of Cambodia, sized by recorded bodies. Panel (b) pools the gap intensity of the exposed cohorts (born 1925–65) across all three censuses; panel (c) repeats the exercise for all other cohorts on the same color scale, as a benchmark for the sex-ratio geography absent the shock.

and with a single measure.

Data. We use *EU&I 2014*, the Europe-wide Voting Advice Application built by the European University Institute for the 2014 European Parliament elections. Respondents reacted to 28 common policy statements (plus two country-specific ones), and the same statements were used to position 242 parties across the EU-28, coded by 28 national teams of political scientists who combined party self-placement with expert opinions. The data is ideal for our purposes for two reasons. First, voters and parties are placed on the same scale in every dimension, so they live in one issue space that is immediately comparable. Second, it lets us run the same analysis across countries with comparable variables.¹⁷ Nonetheless, one should note that respondents self-selected into the samples, so we are careful about drawing conclusions about the general population.

Implementation. From each country’s individual responses we fit a two-dimensional graded-response item-response model, estimated by amortized variational inference, that embeds each respondent’s Likert answers as an ideal point in a latent plane.¹⁸ Parties are then placed by scoring their expert-coded positions through the same fitted model, so voter and party points are in the exact same space. We rotate the plane so that its first axis is the economic left–right dimension, and its second is its orthogonal dimension, whose reflection is anchored on a question on views on the EU, and we standardize both. On this space, we replicate the analysis explained above for the food-desert application, with voters in place of residents and parties in place of grocers: we bin the respondent cloud into hexagonal cells, give each cell a density (the mass of voters it holds) and a distance (to the nearest party), and run the angle-free plane sweep over the rank-normalized (distance, density) plane. A hexagon is ‘underserved’ in the representation sense when it is far from every party *and* dense with voters. A bounded cluster of such cells, encircled by served voters or low density cells, is an H_1 representation desert, and integrating its enclosed membership over the plane gives a per-cell desert score. Note that our aim is not to improve on the estimation of ideal points, to which a substantial literature is devoted (Clinton, Jackman and Rivers, 2004; Martin and Quinn, 2002; Imai, Lo and Olmsted, 2016), but, taking the estimated space as given, to pin down the gaps within it. In other words, assuming the ideal point distribution and the party positions are correct, what we focus on is identifying the correct location and size of the gap.

¹⁷We ran every country available in the data, here we report Germany and Italy as illustrative multiparty cases.

¹⁸“No opinion” answers enter as missing, and a respondent with no substantive answer is dropped.

Results. Figures 13 and 14 show how our pipeline produces visually validated results for Germany and Italy. The *Space* panel (e.g. 13a) places every voter and party in the fitted two-dimensional space, each voter colored by the party she would most likely support and left faint if undecided; it shows the raw joint distribution of the electorate and the party system. The *Density* panel (13b) is the hex-binned voter density alone, or the ideological correspondent of residential density, marking where voters actually concentrate. The *Distance* panel (13c) shades each cell by its distance to the nearest party, similarly to distance-to-grocery, high wherever the party system thins out. Neither density nor distance alone is a gap: an empty fringe of the space is far from parties but holds no one, and a crowded cell next to a party is dense but well served. The *Desert* panel (13d) is the combination the method is built to find: it lights up only where a dense body of voters sits far from every party *and* is enclosed by better-represented voters, leaving represented dense cells flat. Reading the two together, the desert panel isolates the interior of each electorate that the parties collectively fail to reach. And because the score is comparable across cells, it distinguishes the country’s largest representation gap from its minor ones. A quick visual inspection of the parties’ and voters’ relative positions confirms that the representation gaps are found exactly where they should be, conditional on the estimated ideal points.¹⁹ What this gives us is a map of representation gaps as objects in their own right, rather than a single congruence score. Also note that, especially for Italy (Figure 14d), the ‘weird’ shape that representation deserts can have would make them hard to fully detect otherwise.

¹⁹While this visual validation is possible for a few cases, the ultimate goal is to use the method at scale, where the same validation would not be possible.

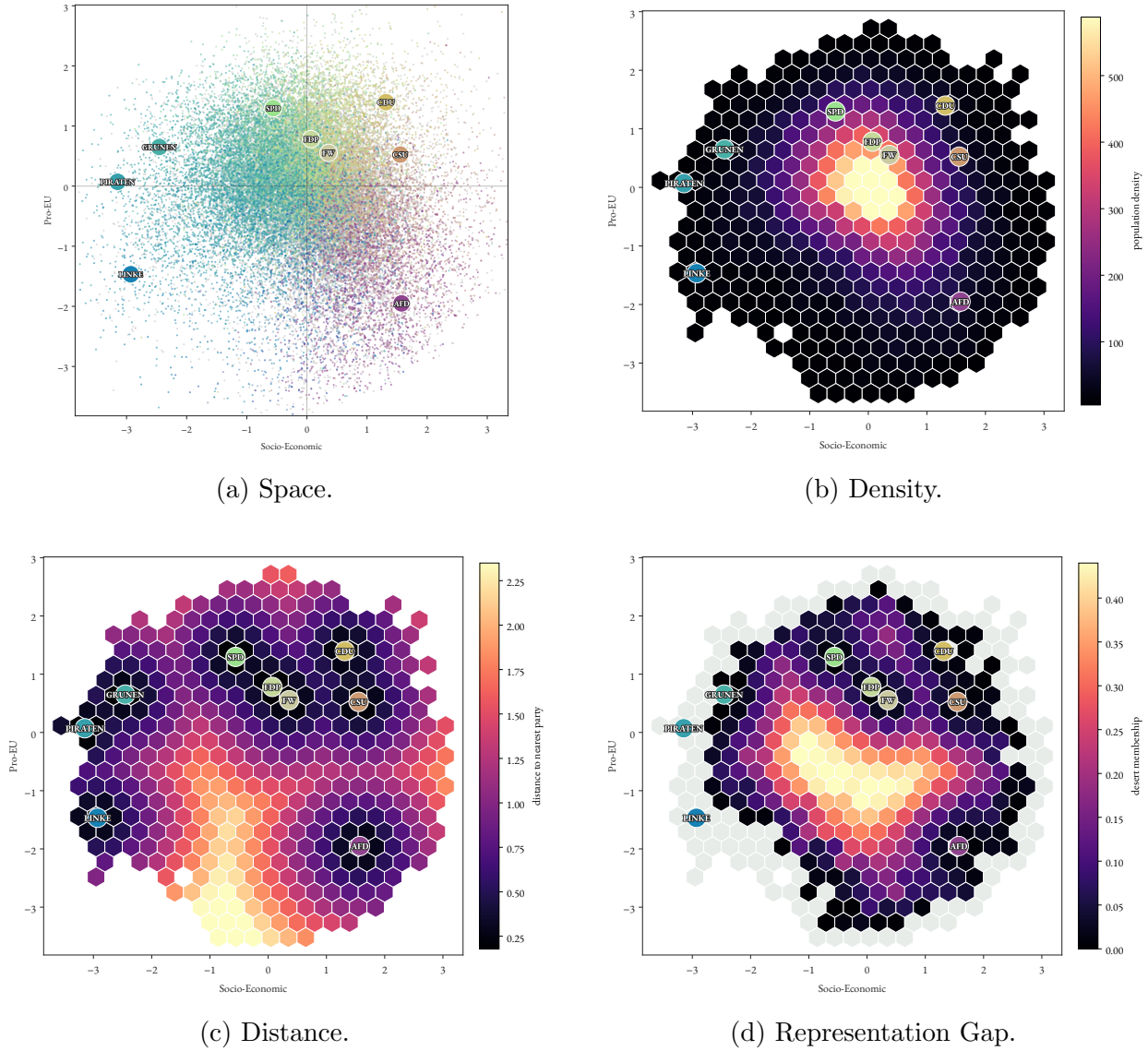


Figure 13: Representation gaps in Germany, 2014. The underlying space is the fitted two-dimensional issue space (economic left–right and a cross-cutting dimension). *Space* (a) places every voter, colored by the party she would most likely support and faint if undecided, and every party. *Density* (b) is the hex-binned voter mass alone; *Distance* (c) shades each cell by its distance to the nearest party. The *Representation Gap* panel (d) lights up only where a dense body of voters sits far from every party *and* is enclosed by better-represented voters.

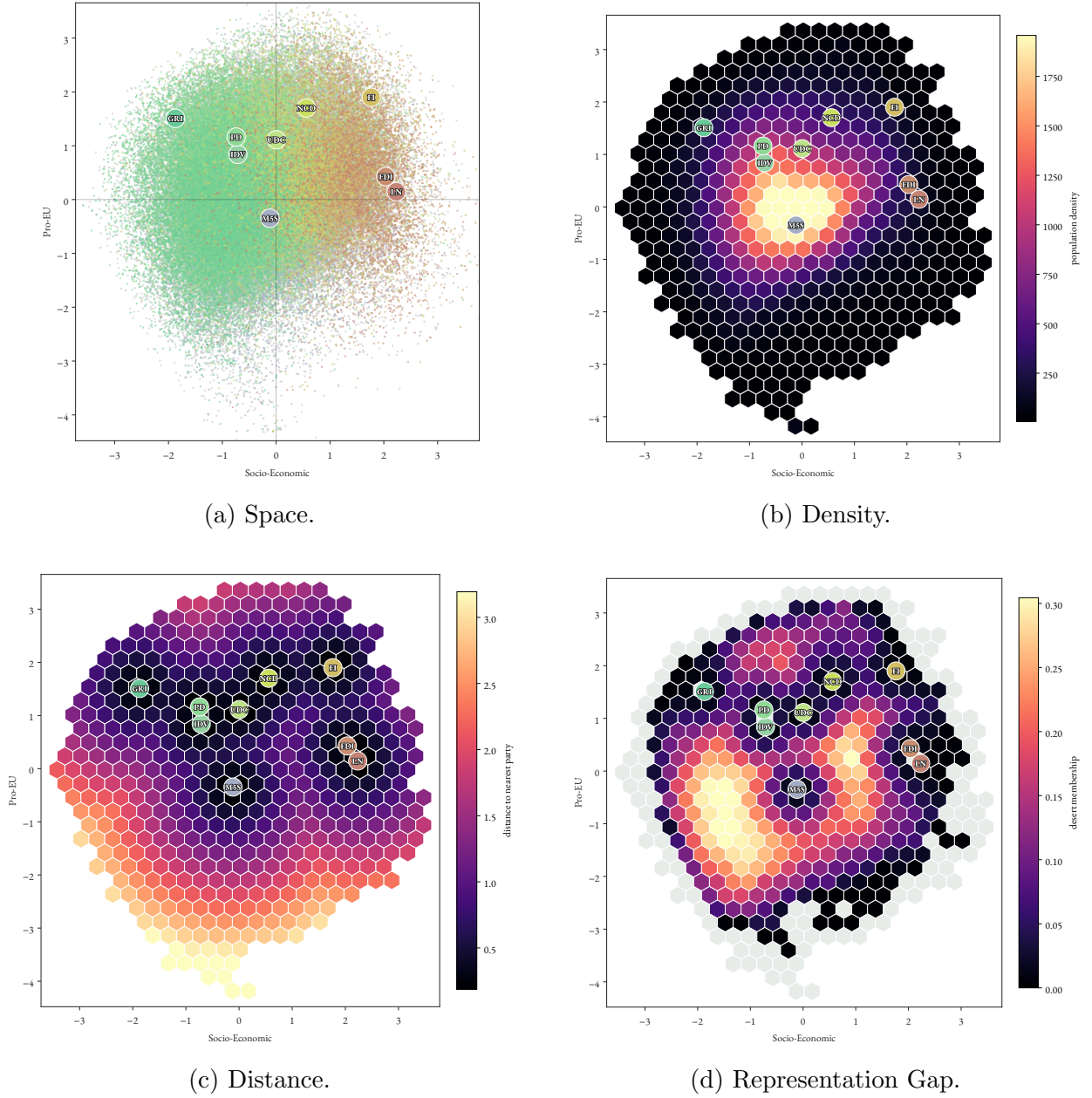


Figure 14: Representation gaps in Italy, 2014. The construction is exactly as in Figure 13: *Space* (a) shows voters (colored by most-likely party, faint if undecided) and parties in the fitted economic-by-cultural plane, *Density* (b) the hex-binned voter mass, and *Distance* (c) each cell's distance to the nearest party. The *Representation Gap* panel (d) isolates the dense, enclosed regions of the electorate that lie far from every party.

5 Conclusion

This paper set out to make a quantity we usually gesture at into one we can measure. We argued that many questions in social science are, at bottom, questions about absence, where resources are missing, who is unrepresented, where a population has been thinned, and that our standard tools, built to summarize the level of a variable defined on pre-drawn units, systematically discard what makes an absence interesting: its arrangement, its individuation, its scale, and its relevance. We therefore defined “nothing” as a quantity of interest in its own right, an enclosed absence with an existence, a location, and a size, and set out the criteria any adequate measure of it should meet. We then proposed one such measure, adapted from topology: treat the data as a shape, and use persistent homology to detect its holes, place them, and rank them by how pronounced they are. Applied to food access, to the demographic imprint of genocide, and to political representation, the same construction recovered gaps that are known to exist but that no existing method sizes and locates systematically, and in each case it agreed with what could be checked independently, from Central Park to the Detroit racial disparity to the Cambodian mass-grave record.

We want to be clear about the division of labor between the problem and our solution to it. The main contribution we hope to make is the reframing: that absence is a first-class estimand, and that posing a question as a nothing-problem is often more revealing than the something-problem. Persistent homology is our answer, and we think a good one, because it reads a gap from the arrangement of the data without being told where to look and reports its structure across all scales at once. But it comes at a cost (a computational one), and it is not the only conceivable answer. We would encourage others to propose rival measures of the same quantity, or improve on the criteria a measure should satisfy.

The three applications are also meant as illustrations, not as the boundary of the method. The construction applies wherever a population is spread over a space and served by a discrete set of providers, and the gaps between them are the object of interest. Access to hospitals, clinics, pharmacies, banks, broadband, polling places, or schools all fit the food-desert template directly, with the served-and-dense conjunction carrying over unchanged. The space need not be geographic. Any latent space in which units and the things that serve them can be jointly embedded, an issue space, an embedding of legislators or bills, a map of topics in public debate, admits the same question of which populated regions no provider reaches.²⁰ Absence in time is equally available: a filtration can sweep a calendar as easily as a distance, so that a gap becomes a period in which something expected fails to appear,

²⁰One should note, however, that the space should be well defined. Word embeddings living in an anisotropic space, for instance, would be not a good use case.

a silence in coverage, a lull in mobilization, a cohort that goes missing. And because the machinery is indifferent to dimension, the same tools extend to absences that are hard to picture, gaps in a high-dimensional space of attributes, that resist visual inspection precisely where a systematic measure is most needed.

Several things remain to be done, and we flag them plainly. The most important is inference. In the current draft, we have only hinted at the resampling and permutation procedures that would be relevant for inference. Implementing those procedures across the applications, so that every gap we report carries a calibrated statement of how confidently it is distinguished from noise, is the natural next step, and one we intend to carry out in future iterations. A second task is robustness: absences are, by construction, defined against what surrounds them, which makes them sensitive to the boundary of the study region and to outliers in a way that presence is not, and we owe a systematic account of how much the located gaps move as those choices vary. A third is substantive follow-through. In the representation case in particular, once data quality is secured, the located gaps are not merely a description but a set of hypotheses, about where new parties can profitably enter, which coalitions the geometry of preferences favors, and which constituencies are durably offered nothing, that the same object is well suited to test.

Regardless of whether the topological route we took proves the best one, we hope the question it answers, how large is the nothing, and where, is one the field will take up.

References

- Achen, Christopher H. 1978. “Measuring Representation.” *American Journal of Political Science* 22(3):475–510.
- Adcock, Robert and David Collier. 2001. “Measurement Validity: A Shared Standard for Qualitative and Quantitative Research.” *American Political Science Review* 95(3):529–546.
- Aktas, Mehmet E., Esra Akbas and Ahmed El Fatmaoui. 2019. “Persistence Homology of Networks: Methods and Applications.” *Applied Network Science* 4:61.
- Allcott, Hunt, Rebecca Diamond, Jean-Pierre Dubé, Jessie Handbury, Ilya Rahkovsky and Molly Schnell. 2019. “Food Deserts and the Causes of Nutritional Inequality.” *The Quarterly Journal of Economics* 134(4):1793–1844.
- Amézquita, Erik J., Michelle Y. Quigley, Tim Ophelders, Elizabeth Munch and Daniel H. Chitwood. 2020. “The Shape of Things to Come: Topological Data Analysis and Biology, from Molecules to Organisms.” *Developmental Dynamics* 249(7):816–833.
- Anscombe, F. J. 1973. “Graphs in Statistical Analysis.” *The American Statistician* 27(1):17–21.
- Apparicio, Philippe, Marie-Soleil Cloutier and Richard Shearmur. 2007. “The case of Montréal’s missing food deserts: Evaluation of accessibility to food supermarkets.” *International Journal of Health Geographics* 6:4.
- Bachrach, Peter and Morton S. Baratz. 1962. “Two Faces of Power.” *American Political Science Review* 56(4):947–952.
- Banister, Judith and Paige Johnson. 1993. After the Nightmare: The Population of Cambodia. In *Genocide and Democracy in Cambodia: The Khmer Rouge, the United Nations and the International Community*, ed. Ben Kiernan. New Haven, CT: Yale University Southeast Asia Studies pp. 65–140.
- Beaulac, Julie, Elizabeth Kristjansson and Steven Cummins. 2009. “A Systematic Review of Food Deserts, 1966–2007.” *Preventing Chronic Disease* 6(3):A105.
- Botnan, Magnus Bakke and Michael Lesnick. 2023. An Introduction to Multiparameter Persistence. In *Representations of Algebras and Related Structures*. EMS Press. arXiv:2203.14289.

- Carlsson, Gunnar. 2009. “Topology and Data.” *Bulletin of the American Mathematical Society* 46(2):255–308.
- Carlsson, Gunnar and Afra Zomorodian. 2009. “The Theory of Multidimensional Persistence.” *Discrete & Computational Geometry* 42(1):71–93.
- Chan, Joseph Minhow, Gunnar Carlsson and Raul Rabadán. 2013. “Topology of Viral Evolution.” *Proceedings of the National Academy of Sciences* 110(46):18566–18571.
- Chazal, Frédéric and Bertrand Michel. 2021. “An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists.” *Frontiers in Artificial Intelligence* 4:667963. Also circulated as arXiv:1710.04019.
- Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. “The Statistical Analysis of Roll Call Data.” *American Political Science Review* 98(2):355–370.
- Cohen-Steiner, David, Herbert Edelsbrunner and John Harer. 2007. “Stability of Persistence Diagrams.” *Discrete & Computational Geometry* 37(1):103–120.
- de Walque, Damien. 2005. “Selective Mortality during the Khmer Rouge Period in Cambodia.” *Population and Development Review* 31(2):351–368.
- de Walque, Damien. 2006. “The socio-demographic legacy of the Khmer Rouge period in Cambodia.” *Population Studies* 60(2):223–231.
- Duchin, Moon, Tom Needham and Thomas Weighill. 2022. “The (Homological) Persistence of Gerrymandering.” *Foundations of Data Science* 4(4):581–622.
- Edelsbrunner, Herbert, David Letscher and Afra Zomorodian. 2002. “Topological Persistence and Simplification.” *Discrete & Computational Geometry* 28(4):511–533.
- Edelsbrunner, Herbert and John L. Harer. 2010. *Computational Topology: An Introduction*. Providence, RI: American Mathematical Society.
- Fasy, Brittany Terese, Fabrizio Lecci, Alessandro Rinaldo, Larry Wasserman, Sivaraman Balakrishnan and Aarti Singh. 2014. “Confidence Sets for Persistence Diagrams.” *The Annals of Statistics* 42(6):2301–2339.
- Feng, Michelle and Mason A. Porter. 2021. “Persistent Homology of Geospatial Data: A Case Study with Voting.” *SIAM Review* 63(1):67–99.

- Fugacci, Ulderico, Sara Scaramuccia, Federico Iuricich and Leila De Floriani. 2016. Persistent Homology: a Step-by-step Introduction for Newcomers. In *Smart Tools and Apps for Graphics (STAG)*. Eurographics Association pp. 1–10.
- Gaikwad, Nikhar, Erin Lin and Noah Zucker. 2023. “Gender After Genocide: How Violence Shapes Long-Term Political Representation.” *World Politics* 75(3):439–481.
- Ghrist, Robert. 2008. “Barcodes: The Persistent Topology of Data.” *Bulletin of the American Mathematical Society* 45(1):61–75.
- Golder, Matt and Jacek Stramski. 2010. “Ideological Congruence and Electoral Institutions.” *American Journal of Political Science* 54(1):90–106.
- Heuveline, Patrick. 1998. “‘Between One and Three Million’: Towards the Demographic Reconstruction of a Decade of Cambodian History (1970–79).” *Population Studies* 52(1):49–65.
- Heuveline, Patrick and Bunnak Poch. 2007. “The Phoenix Population: Demographic Crisis and Rebound in Cambodia.” *Demography* 44(2):405–426.
- Hickok, Abigail, Benjamin Jarman, Michael Johnson, Jiajie Luo and Mason A. Porter. 2024. “Persistent Homology for Resource Coverage: A Case Study of Access to Polling Sites.” *SIAM Review* . arXiv:2206.04834.
- Huber, John D. and G. Bingham Powell Jr. 1994. “Congruence between Citizens and Policymakers in Two Visions of Liberal Democracy.” *World Politics* 46(3):291–326.
- Imai, Kosuke, James Lo and Jonathan Olmsted. 2016. “Fast Estimation of Ideal Points with Massive Data.” *American Political Science Review* 110(4):631–656.
- Kaufman, Aaron R., Gary King and Mayya Komisarchik. 2021. “How to Measure Legislative District Compactness If You Only Know It When You See It.” *American Journal of Political Science* 65(3):533–550.
- Kiernan, Ben. 1996. *The Pol Pot Regime: Race, Power, and Genocide in Cambodia under the Khmer Rouge, 1975–79*. New Haven, CT: Yale University Press.
- Lefkofridi, Zoe, Markus Wagner and Johanna E. Willmann. 2014. “Left-Authoritarians and Policy Representation in Western Europe: Electoral Choice across Ideological Dimensions.” *West European Politics* 37(1):65–90.

- Lum, Pek Y., Gurjeet Singh, Alan Lehman, Tigran Ishkanov, Mikael Vejdemo-Johansson, Muthu Alagappan, John Carlsson and Gunnar Carlsson. 2013. “Extracting insights from the shape of complex data using topology.” *Scientific Reports* 3:1236.
- Luo, Wei and Fahui Wang. 2003. “Measures of Spatial Accessibility to Health Care in a GIS Environment: Synthesis and a Case Study in the Chicago Region.” *Environment and Planning B: Planning and Design* 30(6):865–884.
- Lytle, Leslie A. 2009. “Measuring the Food Environment: State of the Science.” *American Journal of Preventive Medicine* 36(4, Supplement):S134–S144.
- Mahoney, James and Gary Goertz. 2004. “The Possibility Principle: Choosing Negative Cases in Comparative Research.” *American Political Science Review* 98(4):653–669.
- Martin, Andrew D. and Kevin M. Quinn. 2002. “Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999.” *Political Analysis* 10(2):134–153.
- Matejka, Justin and George Fitzmaurice. 2017. Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. pp. 1290–1294.
- McKinnon, Robin A., Jill Reedy, Meredith A. Morrisette, Leslie A. Lytle and Amy L. Yaroch. 2009. “Measures of the Food Environment: A Compilation of the Literature, 1990–2007.” *American Journal of Preventive Medicine* 36(4, Supplement):S124–S133.
- Neupert, Ricardo F. and Virak Prum. 2005. “Cambodia: Reconstructing the Demographic Stab of the Past and Forecasting the Demographic Scar of the Future.” *European Journal of Population* 21(2–3):217–246.
- Nicolau, Monica, Arnold J. Levine and Gunnar Carlsson. 2011. “Topology Based Data Analysis Identifies a Subgroup of Breast Cancers with a Unique Mutational Profile and Excellent Survival.” *Proceedings of the National Academy of Sciences* 108(17):7265–7270.
- Otter, Nina, Mason A. Porter, Ulrike Tillmann, Peter Grindrod and Heather A. Harrington. 2017. “A roadmap for the computation of persistent homology.” *EPJ Data Science* 6(1):17.
- Pitkin, Hanna F. 1967. *The Concept of Representation*. Berkeley, CA: University of California Press.

- Singh, Gurjeet, Facundo Mémoli and Gunnar Carlsson. 2007. Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition. In *Eurographics Symposium on Point-Based Graphics*, ed. M. Botsch, R. Pajarola, B. Chen and M. Zwicker. The Eurographics Association pp. 91–100.
- Skaf, Yara and Reinhard Laubenbacher. 2022. “Topological Data Analysis in Biomedicine: A Review.” *Journal of Biomedical Informatics* 130:104082.
- Sliwinski, Marek. 1995. *Le génocide khmer rouge: une analyse démographique*. Paris: L’Harmattan.
- Sousbie, Thierry. 2011. “The Persistent Cosmic Web and Its Filamentary Structure – I. Theory and Implementation.” *Monthly Notices of the Royal Astronomical Society* 414(1):350–383.
- Walker, Renee E., Christopher R. Keane and Jessica G. Burke. 2010. “Disparities and access to healthy food in the United States: A review of food deserts literature.” *Health & Place* 16(5):876–884.
- Wasserman, Larry. 2018. “Topological Data Analysis.” *Annual Review of Statistics and Its Application* 5:501–532.
- Xia, Kelin and Guo-Wei Wei. 2014. “Persistent Homology Analysis of Protein Structure, Flexibility, and Folding.” *International Journal for Numerical Methods in Biomedical Engineering* 30(8):814–844.
- Zenk, Shannon N., Amy J. Schulz, Barbara A. Israel, Sherman A. James, Shuming Bao and Mark L. Wilson. 2005. “Neighborhood Racial Composition, Neighborhood Poverty, and the Spatial Accessibility of Supermarkets in Metropolitan Detroit.” *American Journal of Public Health* 95(4):660–667.
- Zomorodian, Afra and Gunnar Carlsson. 2004. “Computing Persistent Homology.” *Discrete & Computational Geometry* 33(2):249–274.