Applied Network Science

RESEARCH

Open Access

Identifying loci in mobility networks with applications in New Zealand work commutes: a statistical test for identifying extreme stationary distribution values in Markov transition matrices

Michael J. Kane[1*], Owais Gilani[2,3], Elena Khusainova[4] and Simon Urbanek[5]

*Correspondence:
mjkane@mdanderson.org

[1] Department of Lymphoma and Myeloma, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[2] Mathematics Department, Bucknell University, Lewisburg, PA, USA
[3] Public Health and Community Medicine, Tufts University, Boston, MA, USA
[4] AT&T Labs Research, New York, NY, USA
[5] Statistics Department, The University of Auckland, Auckland, New Zealand

Abstract

Human mobility describes physical patterns of movement of people within a spatial system. Many of these patterns, including daily commuting, are cyclic and quantifiable. These patterns capture physical phenomena tied to processes studied in urban planning, epidemiology, and other social, behavioral, and economic sciences. This paper advances human mobility research by proposing a statistical method for identifying locations that individual move *to and through* at a rate proportionally higher than other locations, using commuting data for the country of New Zealand as a case study. We term these locations *mobility loci* and they capture a global property of communities in which people commute. The method makes use of a directed-graph representation where vertices correspond to locations, and traffic between locations correspond to edge weights. Following a normalization, the graph can be regarded as a Markov chain whose stationary distribution can be calculated. The proposed permutation procedure is then applied to determine which stationary distribution values are larger than what would be expected, given the structure of the directed graph and traffic between locations. The results of this method are evaluated, including a comparison to what is already known about commuting patterns in the area as well as a comparison with similar features.

Keywords: Mobility loci, Directed Graph Clustering, Markov transition matrix

Introduction and background

Patterns of human mobility describe the movement of individuals or the aggregate movement of groups of individuals over time. One class of aggregate movement is *commuting* patterns. That is, the movement of individuals from their home locations to work (assumed to be most often in the morning), as well as their movement from work back home (assumed to be most often in the evening). For a given area and for regular

working days (weekdays, not holidays) these patterns are periodic, they can be quantified, and analyses can be used to understand them.

There is a rich history of research on understanding human mobility patterns from various disciplines including geography, sociology, anthropology, physics, and demography. Barbosa et al. (2018) provide an overview of some earlier approaches as well as recent advances. Research in this domain was reinvigorated in the early 21st century with the increased availability of geo-referenced data at the same time when "big data" was emerging as it's own area of study. A popular source of data for human mobility modeling was based on cell phone telemetry data. Gonzalez et al. (2008) were among the first to make use of individual-level cellular phone data to show that individuals move in regular spatio-temporal cycles that can be modeled. Song et al. (2010) provided both models for human mobility using a Lévy process as well as a characterization for the model, showing a high degree of predictability in the daily patterns of individuals. This development continued with Alessandretti et al. (2020) showing regional, spatial areas in which movement is contained as well as Yuan and Raubal (2016) who explore the distribution of "human activity spaces" and demographic differences in these distributions. More recent work leverages social media data to capture collective movement and predict aggregate movement Abbasi and Alesheikh (2018).

While development of methods for understanding human mobility is interesting on its own, it is also fundamental to areas of research including social science, epidemiology, and infrastructure planning. Epidemiology in particular has recently made use of mobility methods with Gilani et al. (2020) using mobility to both validate existing pollution exposure models as well as identify demographics for which exposure estimates are heavily biased – a question previously posed in Park and Kwan (2017). Other examples, like Alessandretti (2022), Bonaccorsi et al. (2020), Kane and Gilani (2021), Kraemer et al. (2020), highlight the need to incorporate mobility methods into those of existing outbreak models to better evaluate the effect of interventions in the COVID-19 pandemic.

One of the biggest barriers to mobility research has traditionally been acquiring high-quality data. Most mobility data are derived from cell-phone or cell-phone-application data collected by telecommunications companies or large technology companies. Telecommunication data are often proprietary and difficult to obtain. Technology company location data are often purchasable, but are generally biased toward users of the application or owners of devices making the generality of analyses based on these data to the larger public difficult to assess.

Human mobility has been studied using various quantitative approaches including the use of graphs or networks where vertices correspond to spatial regions and edges capture some aspect of movement between those regions (e.g. Hossmann et al. 2011; Ruan et al. 2019; Chen et al. 2022). These graphs can be referred to as *mobility networks* and are distinguished from other types of networks (transportation networks, infrastructure networks, etc.) in that they capture human movement regardless of physical infrastructure facilitating transportation, including roads (Barthélemy 2011). Mobility networks are slightly more general and, because the method presented will work for either, we will prefer the more general term.

Methods for analyzing mobility networks focusing on properties of vertices allows us to determine various characteristics of the corresponding spatial locations.

Vertex properties can be "local" meaning that they can be derived from a vertex and its neighborhood (a subgraph including the vertex). These properties include in-degree, out-degree, whether or not a vertex is a sink, whether or not it is a source, etc. Vertex properties can also be "global" meaning the property is not local. These properties include eigenvector centrality, graph radius, graph diameter, whether or not a vertex is a central point, etc.

In this paper, we propose a new global vertex characteristic of a directed graph identifying "mobility loci" or simply loci, which are vertices that act as "hubs" of movement. Roughly, vertices designated as loci are those that attract movement *to and through* them in greater proportion compared to other vertices. A more precise definition is provided in Section "A permutation test for loci". These loci are identified by accounting for the global properties of the graph in which they reside. This is a statistical property found by estimating a Markov transition matrix from the graph, calculating the stationary distribution of the resulting Markov chain, and then testing to see which vertices have stationary distribution values greater than expected under a null distribution sampled from a permutation procedure.

To showcase this procedure, we consider an application to aggregate commuting patterns for the country of New Zealand based on census data. These data have the advantage that they are relatively unbiased with respect to the populations being sampled (commuters). Spatial groups are defined by Statistical Area 2 (SA2), which partitions the country into areas that are comparable in terms of population and other factors. These data are publicly available making them accessible both through the supplemental materials provided in this paper as well as through the Stats NZ Tatauranga Aotearoa, New Zealand's official government data agency. The process for curating these data and building the mobility graphs could be repeated for other countries providing similar census data.

This paper proceeds by providing a more complete description of the New Zealand commuting data, including a description of how routing data were derived, and a brief visual exploration. Subsequent sections interleave statistical and mathematical concepts with their application to these data and culminate in the procedure of calculating mobility loci alongside their identification for the country of New Zealand. The intention for this format is to both construct a statistical procedure as well as provide insights to better understand commuting patterns in the country. Section "New Zealand commuting data" provides an overview of the data, a description of the preprocessing required to go from raw census data to directed mobility graphs, and an overview of the spatial properties of the resulting mobility graph. Section "The directed mobility graph" characterizes the directed mobility graph and constructs the optimization for finding the stationary distribution. We note that the calculation of stationary distribution is not new and can be found in various sources, mostly online or in the waypoint literature (Navidi and Camp 2004; Hyytia et al. 2006; Mitsche et al. 2014). However, for completeness, we have followed Chang (2007) before providing a formal construction. Section "A permutation test for loci" proposes a test for finding loci, those elements of the stationary distribution that are larger than what is expected when keeping the graph structure fixed and permuting on traffic between SA2 areas, as well as a procedure for finding groups of loci while addressing multiple-testing

challenges. Section "Stationary distribution and loci as global graphproperties" casts the stationary distribution and loci as global vertex properties. Because the notion of mobility loci is a novel vertex property, there is not a direct comparison with other procedures to evaluate the results. However, we do provide analyses quantifying how much information is encoded in these features compared to local graph features including vertex degree and others explained later. Section "Conclusions" includes potential applications for this work, which fit readily into the spatial research framework as well as other potential application areas.

New Zealand commuting data

Data overview

The New Zealand 2018 Census includes, among other questions, information about the main means of travel to work. Based on the answers to this question and respective residence/workplace addresses the Stats NZ Tatauranga Aotearoa, New Zealand's official data agency, publishes a commuter view dataset (Stats 2020) which aggregates the number of usually resident population aged 15 years and over by main means of travel to work. The spatial aggregation is done by "statistical areas" (Stats 2017) which is a spatial partition of the entire country with focus on retaining comparably constant population per partition in urban and suburban areas. The commuter dataset uses partitions induced by Statistical Area 2 (SA2) polygons (represented as shape files) which typically have population of 2000–4000 residents in urban areas.

The data use fixed random rounding to protect confidentiality. Counts of less than 6 are suppressed according to 2018 confidentiality rules (Stats 2019). For the purpose of this analysis we will ignore suppressed values which may lead to a slight under-count in sparsely populated rural areas, but does not affect urban or sub-urban areas. Given our additive treatment of the individual counts we expect the rounding to not have a significant impact given the magnitude of the resulting values.

The census data, as provided, include SA2 of usual residence as well as those of the workplace. They do not include the actual routes taken by individuals on particular days or traffic volume on individual roads, nor the actual time of day that individuals commute to and from work. We assume that the majority of the commuting to work is done during morning hours and commuting to home is done in the late-afternoon/evening hours. Consider the downtown Auckland region, shown in Fig. 1. Figure 2 shows (a) the spatial distribution of commuter residential locations while (b) shows the spatial distribution of commuter work locations, with red indicating the highest counts and white denoting the lowest. These two maps show that residences are evenly dispersed across the greater Auckland area while work locations are relatively concentrated. They also reveal that the spatial distribution for residential and work locations are fairly inverse of one another - high density residential areas have lower work location densities. The yellow-white colored cluster of SA2s on the water in the center-west of Fig. 2a (with small areas) constitutes Auckland's Central Business District, New Zealand's leading financial hub and the centre of the country's economy. The red SA2 in Fig. 2b is part of the Penrose district, an industrial suburb. Unsurprisingly, there are relatively few residents in these commercial areas.

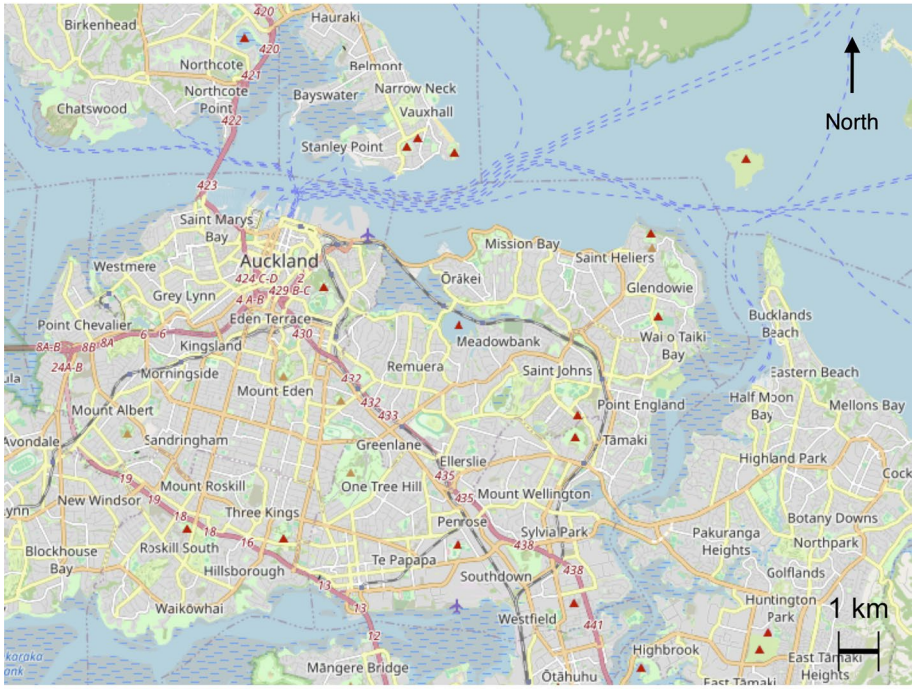


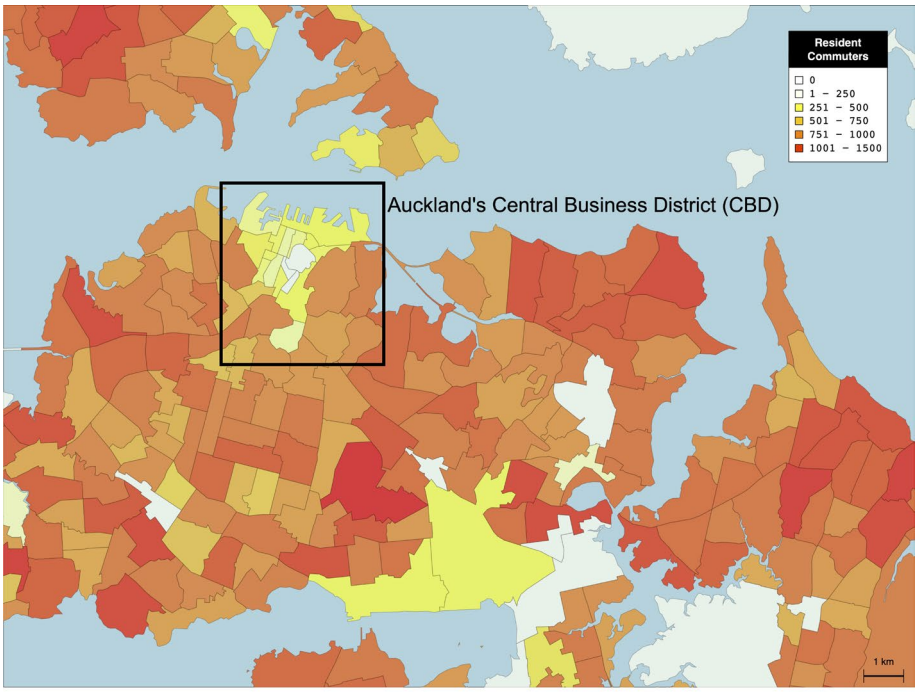
Fig. 1 The downtown Auckland NZ area

Data preprocessing: constructing directed mobility graphs

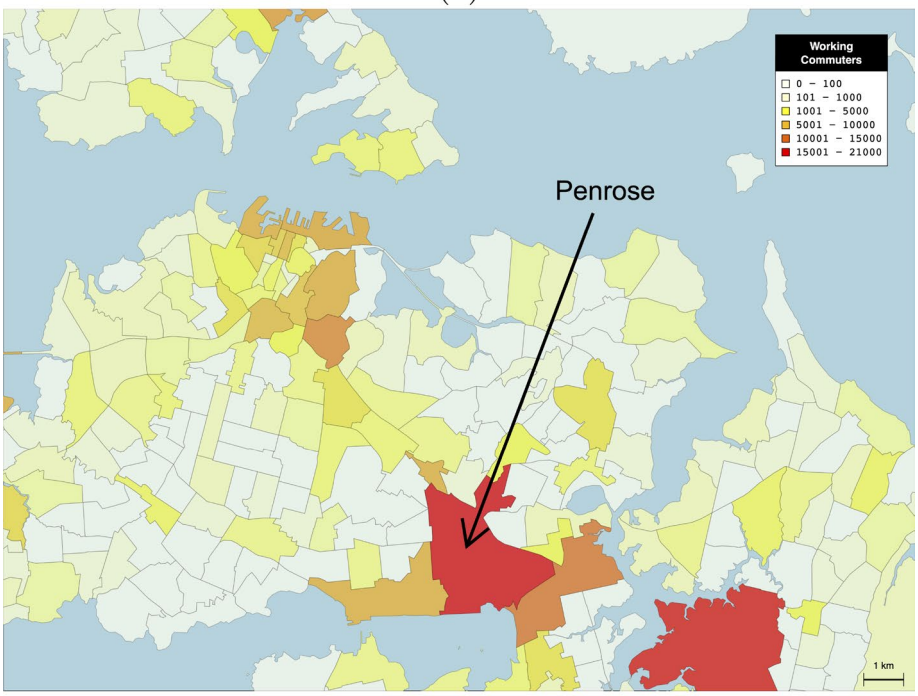
The goal of the preprocessing step is twofold. The first is to determine the most likely route taken by commuters from their residential SA2 area to their work location. We note that on a given day the route taken to work may vary based on many factors including, construction, congestion, etc. and that many residents may not take the most direct or shortest route for any number of reasons such as carpooling, dropping children off at school, etc. However, we assume that the procedure used is sufficient to capture commuting paths for most of the residents enough of the time for the analysis to be valid.

In order to model actual movement through space, we focus on the commutes. The GraphHopper routing engine (GraphHopper 2022) was used to infer the most likely route from home to work, and road topology was sourced from Open Street Map (OpenStreetMap 2022). The resulting route is represented as a sequence of line segments of the most likely roads taken to work/home SA2 area, which is then used to infer the sequence of SA2s along the way using a spatial join. One additional complication is the fact that SA2s often use roads as boundaries so small deviations may cause apparent frequent movement between adjacent SA2s along the road. To counteract that effect we use 10 m buffers around borders and will consider a transition from one SA2 to another only if the route has fully left the SA2 including the extra margin. Sequences can be reversed for work-to-home commutes and the same approach can be used for different modes of transport (public transit, cycling, etc.).

The second goal of the preprocessing step is to aggregate the individual-level commuting sequences into a directed graph. SA2 areas are represented as vertices in the graph. Any movement from one SA2 to another constitutes a directed edge. The weight of the



(a)



(b)

Fig. 2 **a** Spatial distribution of commuter residential locations and **b** Spatial distribution of commuter work locations in downtown Auckland

edge is the count of transitions between pairs of SA2s. Note that by design, edges can only lead from one SA2 to its neighboring SA2s.

The subset of the resulting graph corresponding to the region around central Auckland is illustrated in Fig. 3. The graph represents the movement from home to work, typically corresponding to the home-to-work commute. Each arrow represents the directed movement from one SA2 to another. In many cases there is a movement in both directions for a pair of areas, where some people leave one area and others enter that area. In order to visually distinguish the magnitudes of the flows, we define a "major" direction which, for a pair of vertices, is the edge with the higher weight. In Fig. 3 we color the major direction edges in red and correspondingly the opposite minor direction (if it exists) in blue.

The total width of the red arrows shows us the general major movement between regions. Movements that are counter to the major direction are then seen in blue. We can see an entire series of such counter-movement in the center of the plot, corresponding to the main artery in the region: State Highway 1. The presence of those large counter-movements should be noted because it contradicts a common assumption that workers move from suburbs into the city for work in a "spoke-and-hub" fashion. This is not supported by the visualization. There is a significant north–south flow in both directions, while east–west flows dominate in the direction towards the city.

Figure 4 shows the distribution of the in-flows with red indicating the highest counts and white denoting the lowest. In-flows count all individuals who enter an SA2,

Fig. 3 Commuter movement between SA2 areas at commute time from home to work in the central Auckland region. Edges are drawn from the area centroid. The thickness of the directed edges is proportional to the square root of the number of people moving from one area to another, while the arrows indicate the direction. The major direction of movement has been drawn in red and minor in blue, therefore edges with significant blue thickness have relatively high counter-movement against the major direction

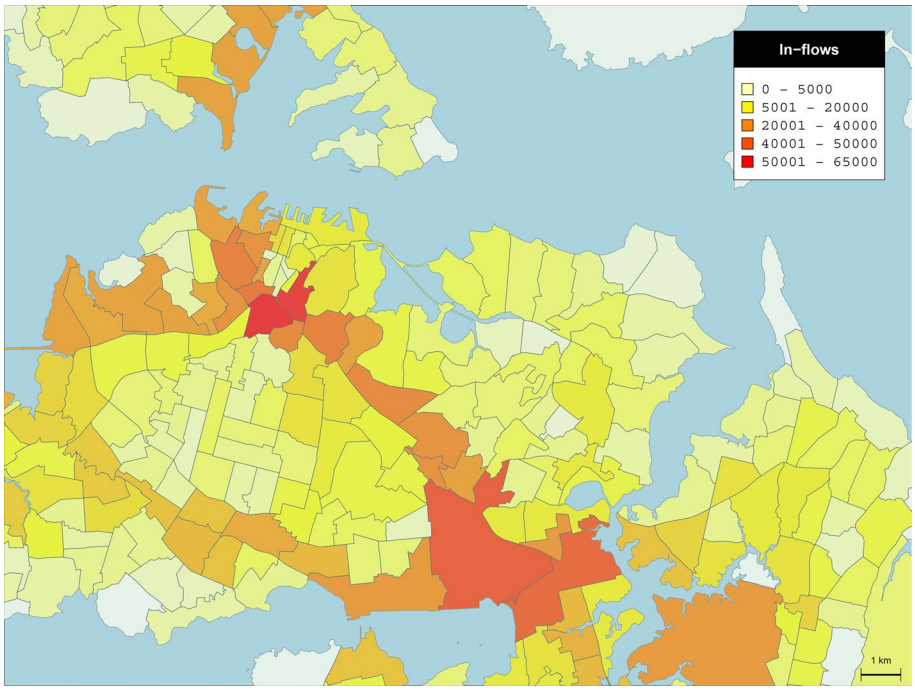


Fig. 4 Distribution of the traffic in-flows, the count of all individuals who enter an SA2, regardless of whether they stay there or subsequently leave

regardless of whether they stay there or subsequently leave. The figure reveals that, as expected, contiguous areas with high traffic are associated with major highway corridors.

The directed mobility graph

The result of the preprocessing step on the home-to-work commute data is the construction of a *directed mobility graph* that encodes the SA2s as vertices and the aggregate movement between adjacent spatial areas as edge weights in a directed graph. This can be thought of as an extension to the origin–destination matrix described in de Dios Ortúzar and Willumsen (2002), containing not only start and endpoints but the intermediate transitions as well. We then partition the directed graph into strong components, i.e., sets of vertices that can be reached from any other vertex in the component. Figure 5 shows these strong components. The largest one, in red (component 1), contains 1426 SA2 areas covering most of the North Island and has the largest population, and will be the focus for the rest of this analysis.

Histograms of the edge weights in the largest component appear in Fig. 6. The weights appear to decrease exponentially, indicating few transitions between most adjacent SA2 areas, except for a few. This is likely because a large portion of the SA2 areas correspond to rural areas, where there are fewer transitions in general as well as those SA2 transitions in more suburban and urban areas with populations who are not commuting to high population-density areas, like city centers.

Figure 7a shows histograms of the in- and out-degrees vertices in the largest component. These two distributions are relatively symmetric with the out-degree histogram having a slightly larger mode and the in-degree histogram having a slightly heavier tail.

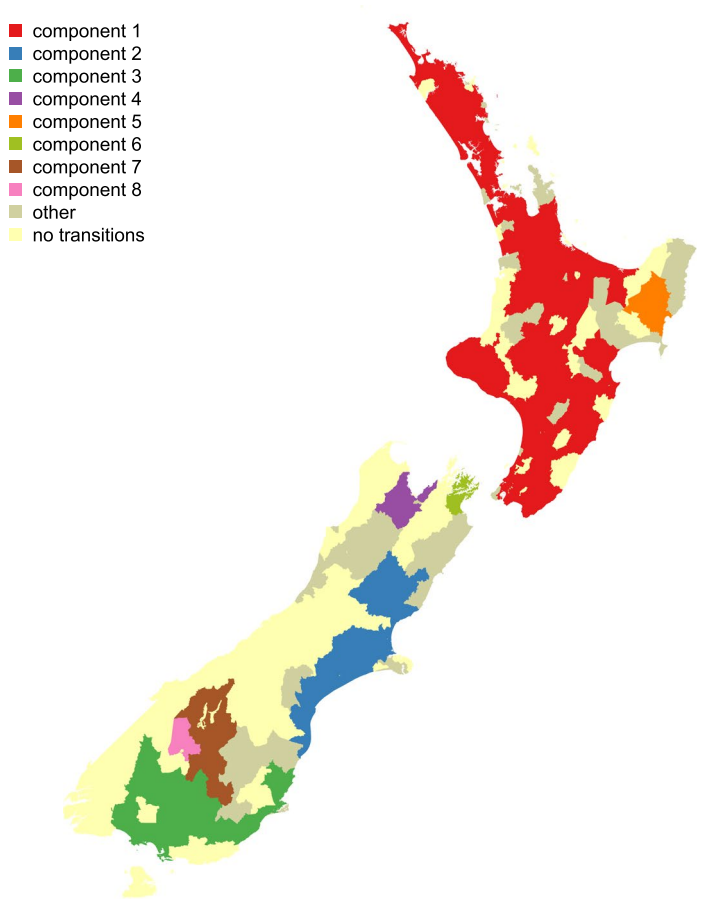


Fig. 5 Strong components in the New Zealand commuter mobility graph for the home-to-work commute

Since these are taken from a strong component and each vertex must have at least one in-edge and one out-edge we may conclude that, on average, vertices incoming and outgoing traffic is balanced. Figure 7b shows a histogram of the log of the edge-weighted in- and out-degrees. The out-degree distribution is sightly more concentrated than that of the in-degree. This is likely an artifact of the difference between rural/suburban and urban SA2 areas, where in rural/suburban areas, a greater proportion of commuters travel in the direction of more urban SA2 areas and have a larger weighted out-degree. Urban areas tend to have more symmetric weighted in- and out-degree values.

Directed mobility graph to its stationary distribution

The directed mobility graph with n vertices can be represented as a matrix, $M \in \mathbb{R}^{n\times n}$, quantifying the aggregate movement between adjacent areas with the rows corresponding to the "from" (origin) location, columns corresponding to the "to" (destination) locations, and elements of the matrix corresponding to the number of people moving between respective locations. The sum of a column captures the total movement of people to a location.

Let $P \in \mathbb{R}^{n\times n}$ be the matrix that results from normalizing over the rows of the mobility graph and let $P_{i,\cdot}$ denote the ith row of P.

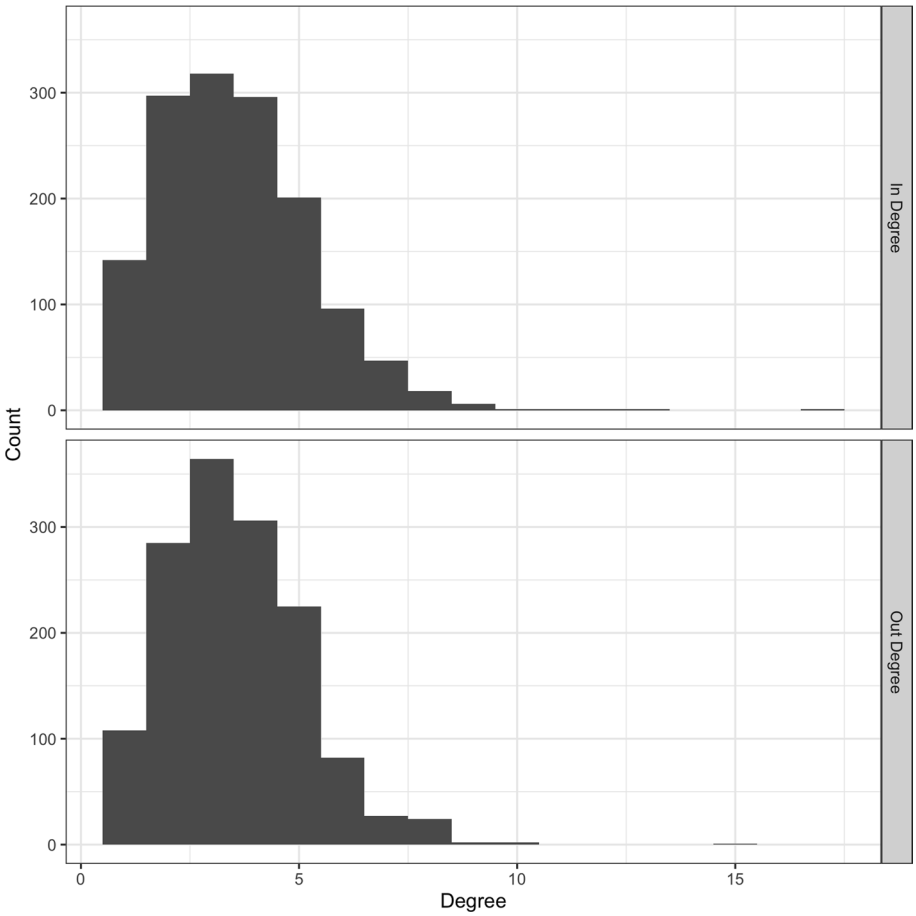


Fig. 6 The histogram of in- and out-degree of the largest strong component of the mobility graph

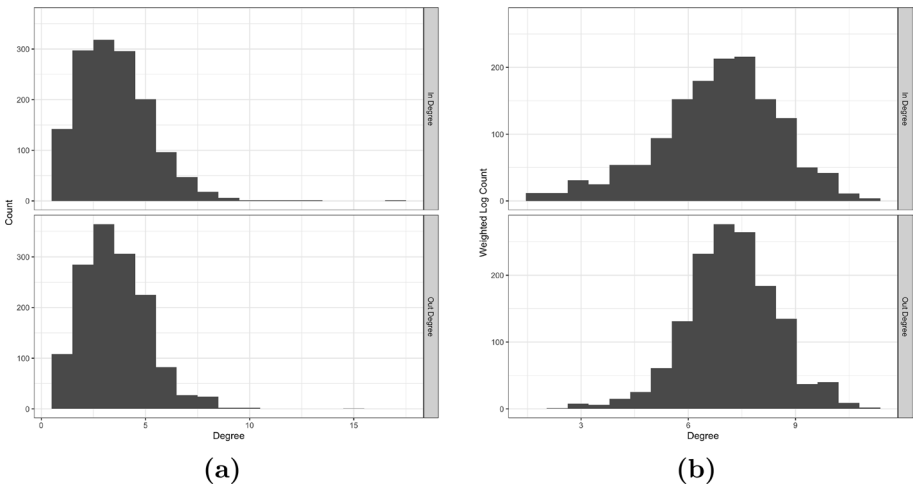


Fig. 7 **a** The histogram of in- and out-degrees for home-to-work commutes of the largest strong component. **b** The histogram of log of edge-weighted in- and out-degrees for home-to-work commutes of the largest strong component

$$P_{i,\cdot} = M_{i,\cdot} \Big/ \sum_{j=1}^{n} M_{i,j}$$

The scaling provides a mobility measure relative to individuals, rather than total movement between SA2s, allowing us to directly compare movement patterns between areas with different population densities. It is also equivalent to estimating the transition probabilities between SA2 areas using the maximum likelihood estimator, under the first-order Markov assumptions. This matrix will be referred to as the Probability Transition Matrix (PTM).

Figure 8 shows the histogram of the non-zero transition probabilities for the largest strong component in the New Zealand data for both the home-to-work commute as well as work-to-home. There are several things to note. First, the home-to-work mobility matrix is the transpose of the work-to-home mobility matrix. However, the same relationship does not hold for the probability transition matrices since they are normalized by the total movement by row. Second, for the home-to-work histogram counts drop off roughly exponentially. This may be because a large number of people live in suburban areas, outside of the city. The first few transitions for these commuters can vary drastically, depending on their destination, corresponding to low-probability transitions. As they transition to secondary or primary roads, "traffic flows" become more regular with less variation on commuting patterns, corresponding to higher transition probabilities on fewer roads. Third, there is a spike at transitions near a value of one. This likely corresponds to traffic along highways outside of urban areas where individuals are commuting toward urban areas.

Calculating the stationary distribution

The PTM is an object allowing us to create a procedure for evaluating the aggregate movement of individuals in between SA2s as a Markov Chain. To quantify mobility with these data, we propose a feature based on the stationary distribution of SA2 areas, which is defined as the probability distribution

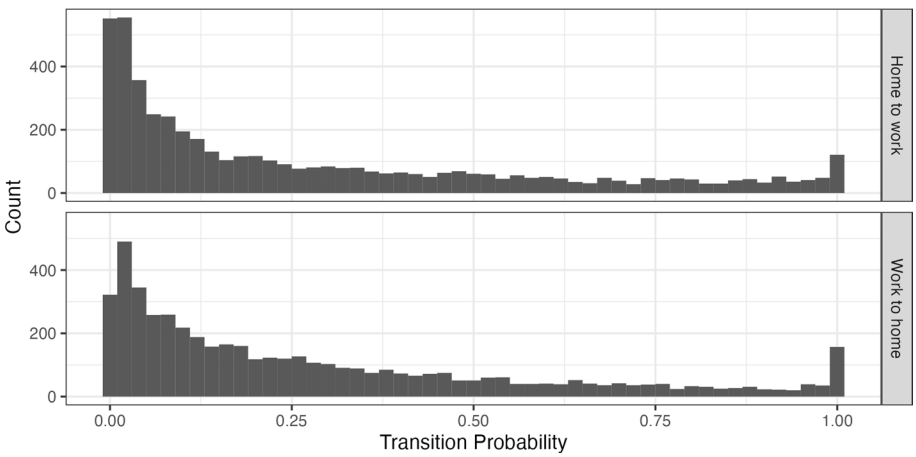


Fig. 8 The histogram of values of the transition probabilities for "Home to work" and "Work to home" commutes

$$\pi = P\pi, \tag{1}$$

where P is the PTM. In order to be able to calculate the stationary distribution, and for it to be unique, two conditions on P must be satisfied (Chang 2007). First, the matrix P must be *irreducible*, meaning that there is a path from any area to any other area in the mobility graph. Second, P must be *aperiodic*, meaning that the greatest common divisor (gcd) of $\{m : (P^m)_{i,i} > 0\}$ for each $i \in \{1, ..., n\}$.

The first is satisfied by conditioning on strong components in the mobility graph, which ensures that there is a path from each vertex to any other vertex in the strong component. The second can be checked directly by examining the diagonals of P^k for $1 \leq k \leq n$ since the strong component condition guarantees that the path length from a vertex to itself has length less than or equal to the number of vertices in the strong component. The gcd can then be calculated for each of the diagonal elements of P^k. If the gcd of any of those values is not 1, then the PTM is periodic and convergence is not guaranteed.

When these two conditions are met, we can solve for π directly by first rearranging the terms

$$\begin{aligned} 0 &= P\pi - \pi \\ 0 &= P\pi - I\pi \\ 0 &= (P - I)\pi \end{aligned}$$

for the identity matrix $I \in \mathbb{R}^{n \times n}$. We now have a linear system of n equations of n variables. To constrain π so that its sum is one, we add the following row to the system.

$$\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} (P - I)\pi \\ \pi \end{bmatrix}$$

where the bracket signifies the matrix resulting from adding new rows (1 to the left side and the vector π on the right) to the system of equations, plus the constraints that the value of π must be at least zero and at most one, define a constrained, linear-optimization problem whose solution can be found via standard methods.

There are two interpretations for the stationary distribution in this setting. The first, standard statistical interpretation is that if an individual were to start at a random SA2 and proceed to an adjacent SA2 according to the probabilities in the PTM, then the stationary distribution is the proportion of time the individual visits each SA2 area as time goes to infinity. Another related interpretation is that the chain is a random dynamical system, with discrete time and discrete state space. The stationary distribution π is a *fixed-point attractor* since it is mapped to itself by the PTM as described in Eq. 1. Roughly, we can interpret SA2s with relatively large stationary distribution values as areas individuals tend to *go to and through*, regardless of their starting point.

A permutation test for loci

The attractor, as defined in this paper, is π, the vector of stationary distribution values with index corresponding to SA2 area. The attractor's values depend on two properties of the mobility graph. First is the graph structure, which is characterized by the connectivity between vertices, encoded with the edges. Second is the individual-level preference

for direction, which is encoded by the edge weights as transition probabilities. The stationary distribution of SA2's show that some elements have values much larger than others. However, these large observed values might be unremarkable given the structure of the graph. Based on this observation a natural question to ask is, "which attractor elements have large values that are greater than expected, *given the structure of the mobility graph for New Zealand's largest North Island strong component*?" Posing the question in this way allows us to identify those SA2 areas that individuals tend to go to and through, independent of the overall population of the area or the total number of people going to or through an area, at a rate that is unusually higher than expected. We term such an SA2 area a *locus*.

Samples from the null distribution of attractors can be derived by randomly permuting the edge weights of the mobility graph, and then calculating the stationary distribution. Fixing the structure of the graph tailors the distribution to that of the mobility graph, while permuting over edge weights fixes the *total movement* captured by the mobility graph. We will refer to the null attractor distribution as Π. Let Π_i and π_i denote the i^{th} element of Π and π respectively, and let α be a suitable cutoff. Then, we formally define the element π_i as a *locus* if

$$\mathbb{P}\{\Pi_i \geq \pi_i\} \leq \alpha. \tag{2}$$

Figure 9 shows the stationary distribution, in black, of the "home to work" and "work to home" commmutes on the original and log scale along with the quantile values of Π, in red, when $\alpha = 5\%$. The values are in descending order of the "home to work" commute stationary distribution. The largest stationary distribution element start with a value of 0.133 and drops off quickly to values close to zero. The corresponding quantile values are relatively small meaning that it is unlikely that larger stationary distribution values would occur under Π. The graphs of the work-to-home commutes are quite different. The stationary distribution values are all very close to 0.000701, which is the expected value of the probability mass function of the uniform distribution on the counting numbers from one to 1426 (the number of vertices in the strong component). A total of 87 of the stationary distribution values fall below the 5% threshold.

Figure 10 shows the individual estimates of the tail-probabilities (p-values) for home-to-work and work-to-home commutes. Together, the plots suggest that home-to-work commutes see the concentration of individuals from their home locations to areas with higher than expected stationary distribution values (*loci*), while the work-to-home commute does not experience this same phenomenon and behaves more closely to uniform mixing over the strong component.

A procedure for identifying high-probability SA2 areas for home-to-work commutes

The previous section essentially amounts to a statistical test where elements of π are used to calculate the tail probability (p-value) under the null distribution, which was sampled by a permutation procedure. While this procedure is effective for testing individual elements of π, identifying sets of loci brings with it the multiple testing challenge. This challenge is mitigated by the fact that, as stated previously, we are

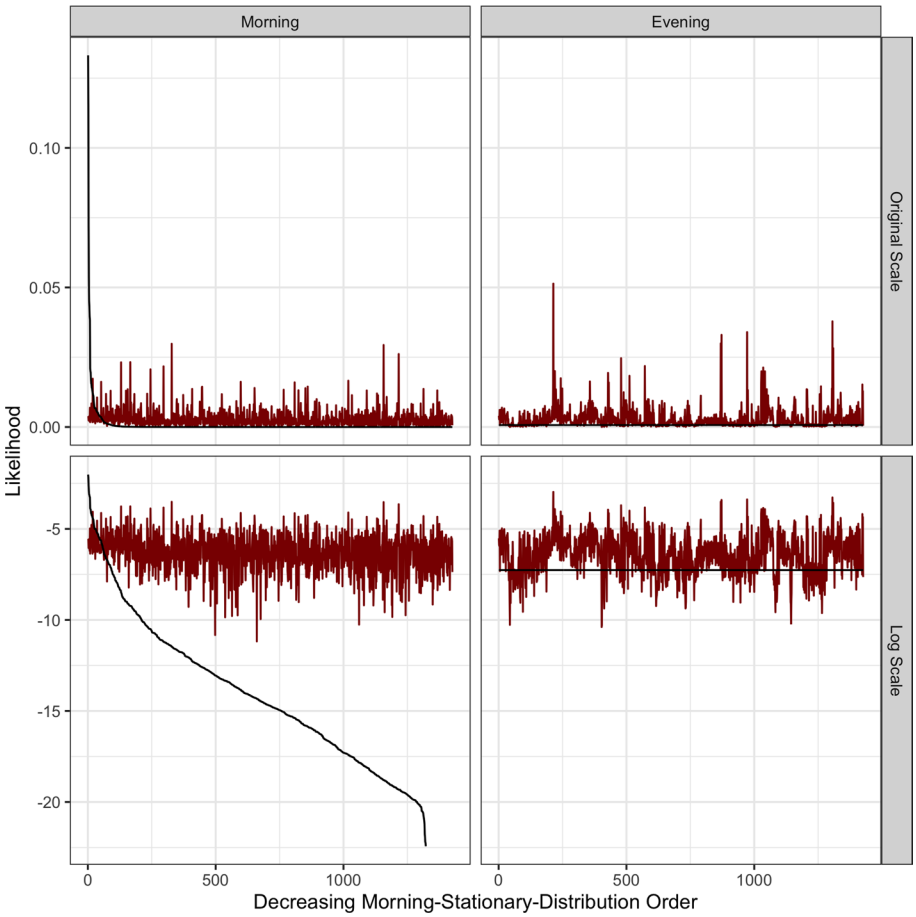


Fig. 9 The stationary distribution of all vertices, in black, in descending order of the "home to work" stationary distribution, of the "home to work" and "work to home" commutes on the original and log scale along with the quantile values of Π, in red, when $\alpha = 5\%$

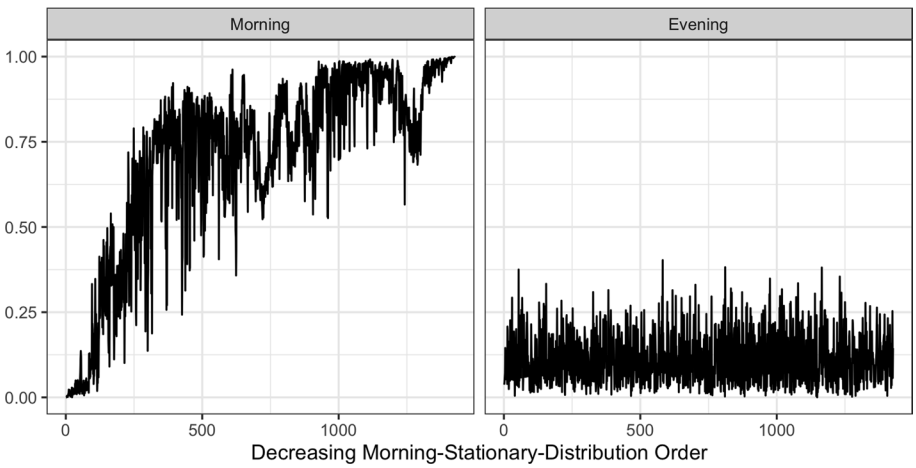


Fig. 10 The tail probabilities and stationary distribution of all SA2's, ordered by largest to smallest home-to-work stationary distribution. Home-to-work commute was assumed to take place in the morning for all commuters, while work-to-home commute was assumed to happen in the evening

interested in the intersection of SA2 areas whose stationary distribution is large *and* whose value is greater than expected. For the rest of this section we will focus on home-to-work commutes.

Let $\Pi_{(i)}$ and $\pi_{(i)}$ be the *i th* values of Π and π ordered in decreasing value of π. Let

$$X_{(i)} = \mathbb{P}\{\Pi_{(i)} \geq \pi_{(i)}\} \tag{3}$$

be the tail-probability of $\pi_{(i)}$ with respect to $\Pi_{(i)}$. Let k be an integer from 1 to n, the number of vertices in the strong component, and let

$$X_{(1:k)} = X_{(1)}, X_{(2)}, ..., X_{(k)} \tag{4}$$

be the vector of tail probabilities corresponding to the k largest stationary distribution value. This can be thought of as a vector of p-values that can be adjusted for using any of the standard methods for multiple comparisons (Bonferroni 1935; Benjamini and Hochberg 1995; Benjamini and Yekutieli 2001; Hommel 1988, etc.) In this analysis we will show results based on Benjamini-Hochberg procedure. This is the least conservative of the listed methods and the liberalness of the procedure is likely warranted because, as will be shown later, the tests are not independent, with sets of loci tending to be highly spatially correlated.

To find the set of loci for the mobility graph, we find

$$\arg\max_{k} \sum_{i=1}^{k} \{X_{(i)} < \alpha\}, \tag{5}$$

the k that maximizes the number of loci. We then report those SA2 areas with adjusted tail probabilities less than α. We do note that by adjusting over different vector sizes, we are essentially doing a "test-of-tests." However, we remind the reader that the goal of this procedure is not to get the adjusted p-values, but rather it is to maximize the number of loci while taking into account the multiple testing problem for each value of k.

Figure 11 shows the number of significant loci, up to k, after adjusting the tail probability ($\alpha = 0.05$) according to the Benjamini-Hochberg procedure. Roughly, as the number of the largest stationary distribution elements increases, so does the number of significant loci. This continues until new stationary distribution values are not enough to justify the increased number of tests and the number of significant loci

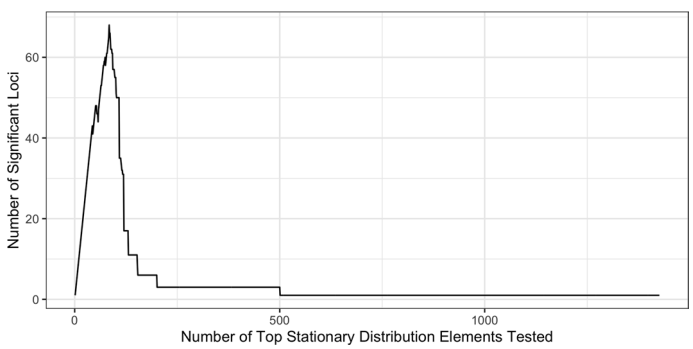


Fig. 11 The number of significant loci conditioned on top stationary distribution values

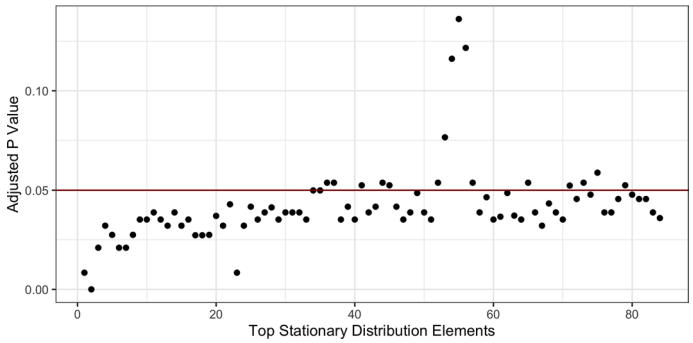


Fig. 12 The False-Detection-Rate (FDR) adjusted *p*-values

begins to decrease. However, the increase is not strict and the graph has three local maxima.

The value of k maximizing Eq. 5 for the largest strong component in the New Zealand mobility graph (with $\alpha = 0.05$) is 68. The graph of the adjusted p-values is shown in Fig. 12. From the graph we can see that it is not true that all of the top stationary distribution values are statistically significant. Depending on the underlying graph structure, we may be able to add more of the top stationary distribution vertices, even when some are not significant, to find the maximum number of adjusted, significant vertices.

Figure 14 is a spatial visualization of the loci in the largest strong component for New Zealand. Loci appear in central Auckland, Wellington, Palmerston North, Hamilton, Kerikeri, Kaitāia, and Rotorua. Local spatial representation of these loci is shown in Fig. 13 (Rotorua is not pictured because it is a singelton). Although the detection of loci in major cities is intuitively expected, even more rural towns like Kerikeri and Kaitāia with population of only around 8000 and 6300, respectively, are hosts of loci which is consistent with their role as hubs in a more sparsely populated region. We note that it is not sufficient for there to only be a few roads in a rural area to have a loci. It could be the case that those roads are difficult to access, in which case its stationary distribution may be small (it would not attract movement). The loci can be interpreted as places individuals are commuting *to and through* in greater proportion compared to other areas, and may be independent of total traffic moving through an SA2 area. In general, delays at loci likely translates to proportionally greater systemic delays when movement through these areas is hampered.

Stationary distribution and loci as global graph properties

The stationary distribution of the strong component as well as the loci, as formulated here, describe global properties of the mobility graph. Intuitively, this can be seen from Section "Calculating the stationary distribution" where the linear-optimization is over the probability transition matrix of the entire graph. The loci feature inherits the global property characterization since it is a subset of the stationary distribution, which is global. These properties can also be seen as properties of individual, corresponding vertices: stationary distribution (proportion of times individuals go to and through that vertex) and loci status (yes or no).

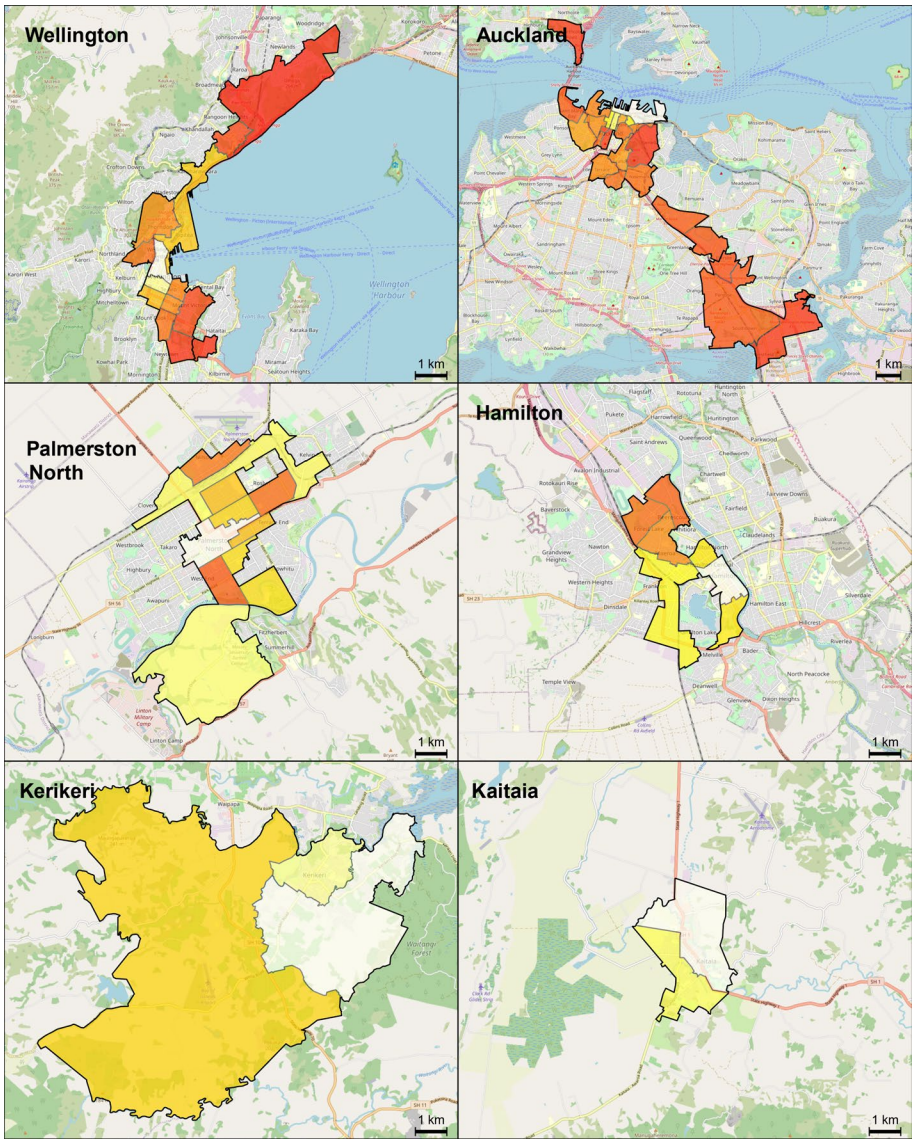


Fig. 13 Loci in the various regions of the North Island, colored by stationary distribution values (yellow corresponds to larger stationary distribution values)

The procedures for calculating both the stationary distribution of the mobility graph and the loci encode novel feature information in the mobility literature and are not represented by existing supervised or unsupervised procedures. As a result, a direct comparison between it and similar procedures is not possible. However, it remains important to evaluate these features in terms that allow us to distinguish them from other related approaches and provide a more intuitive understanding of the mobility information that they encode. To do this, we assess how much information is encoded in this feature compared to four standard *local* features that can be classified as either local properties of vertices or features that are calculated from the user trajectories, the sequence of SA2s defining a commute in this paper. The first local graph vertex property is the *in-degree* of each vertex, i.e., the number of edges

Fig. 14 Loci in the North Island strong component

going from neighbors of an SA2 area into the SA2 area. The second is the *weighted in-degree*, which multiplies each in-edge by the number of commuters coming into the SA2 area. The first trajectory feature will be termed the *total incoming traffic*. It counts the total number of commuters going *to* (and not *through*) an SA2 area from any location. It is equivalent to the weighted in-degree of the source-destination network. The fourth and final is the *total traffic* which measures how many commuters start at, go through, or end at a given SA2 area. This is calculated by taking all of the trajectories and counting how many include a given SA2. Together, these four features will be referred to as the *comparison features*.

This section provides three results supporting the claim that both the stationary distribution as well as the loci provide distinct and potentially valuable information when understanding population-scale human mobility. First, up to approximately one-quarter of the variation in the stationary distribution is captured by the trajectory features. This implies that the global features encoded in the stationary distribution are partially captured by information in individual trajectories. Second, because most of the variation cannot be explained in terms of comparison features, the global features capture information that is distinct from them. This information is likely from the graph structure, which cannot be recovered from the local graph or trajectory features presented. Third, when the trajectory features are regressed onto the stationary

Table 1 Association between local graph features and stationary distribution

Metric	*p*-value	Linear adj. R^2	RF OOB R^2.
In-degree	0.005	0.005	0.001
Weighted in-degree	0.561	≤ 0.001	≤ 0.001
Total incoming	< 0.001	0.143	0.150
Total traffic	< 0.001	0.249	≤ 0.001

Linear adj. R2 is the adjusted R^2 value of the linear regression. RF OOB R^2 is the out-of-bag R^2 value of the random forest model.xx'

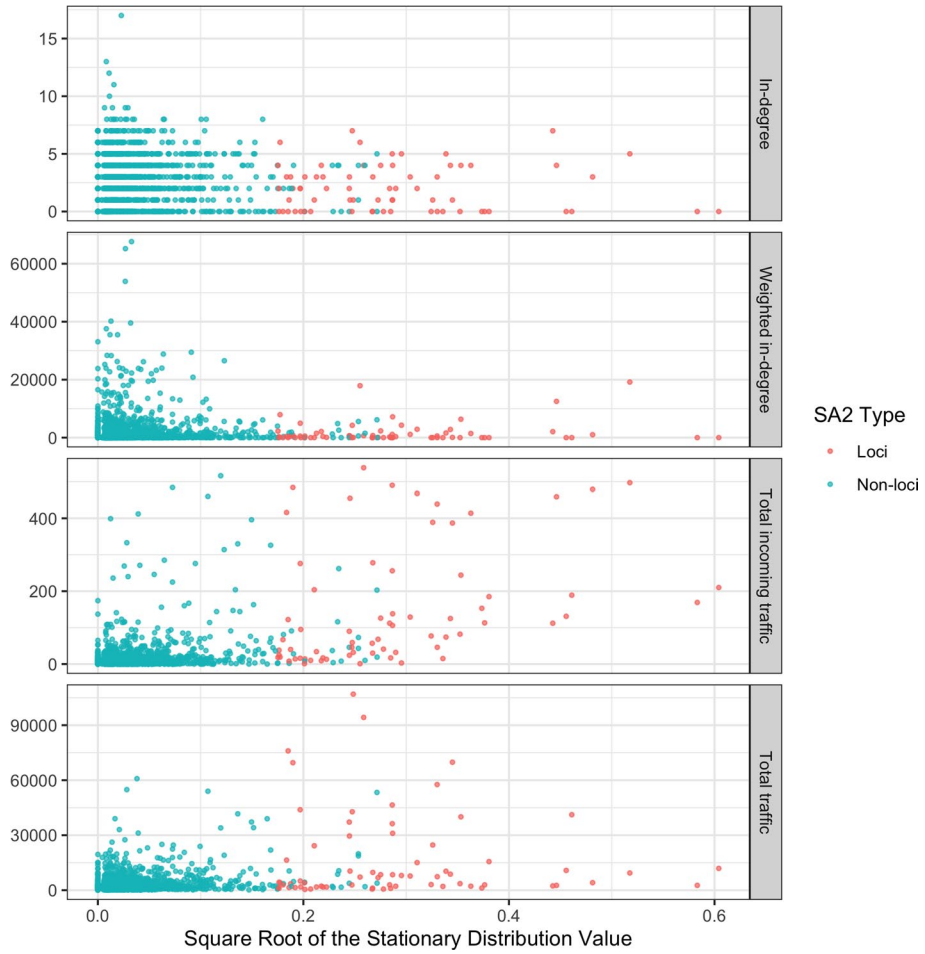


Fig. 15 Scatter plots of the local features versus the square root of the stationary distribution

distribution, the loci have larger corresponding residual values thereby reinforcing the claim that loci identify vertices where there is more than expected mobility.

Admittedly, the fact that these four local features fail to encode the global features being proposed does not imply that there are not other features that do. There are not general theorems telling us under which conditions global properties can be recovered from the features we are comparing against. The claim that they cannot in this case, is left as a conjecture.

Figure 15 shows a scatter plot of the the square root of the stationary distribution (x-axis) versus each of the comparison features (y-axis), along with the loci status of each SA2 area. The graph shows that loci tend to have higher stationary distribution values but not necessarily higher comparison-feature values. Table 1 shows the associative relationships between each of the comparison features and the stationary distribution. The first column is the p-value of the slope coefficient of the linear model regressing the comparison feature onto the stationary distribution. The second column gives the adjusted R^2 values of the associated linear models. The third column gives the out-of-bag R^2 values of the associated random forest model. The third column is provided to show that the associative information exists on a linear subspace and can be well-characterized using a linear model. Other linear model regression diagnostics were performed, including a QQ-plot of the residuals, which showed that the residuals were approximately normal but did contain heavy, symmetric tails. The table shows that, except for the "Weighted in-degree" variable, each of the comparison features had significant linear association with the stationary distribution. However, the comparison features were not able to account for a large portion of the variation with the "Total traffic" variable accounting for 24.9% and others accounting for less.

Since Table 1 shows that "Total incoming" and "Total traffic" are associated with the stationary distribution, similar linear and non-linear models were constructed to quantify the associative relationship between these variables. Table 2 shows the relationship between the two trajectory features ("Total incoming" and "Total traffic"); and, the relationship between the stationary distribution and the two trajectory features together. The "Total incoming" variable explains 24.9% of the in-sample variation of the stationary distribution using a linear model and 14.4% of the out-of-sample variation using a random forest model, while "Total traffic" variable explains 14.3% and 0% of the variation, respectively (Table 1). "Total traffic" accounts for 33.4% and 19.3% of the variation in "Total incoming traffic", and the combined trajectory features account for 26.1% and 23.0% of the variation in the stationary distribution, respectively (Table 2). From this we can conclude that although some of the variation in the stationary distribution can be explained by the trajectory features, it is only about one quarter of the stationary distribution's total variation.

Figure 16 plots the fitted vs. residuals values for the combined linear model on the left and the regression between the trajectory features on the right. From the plot on the left we can see that the loci (colored in red) tend to have large, positive residual values. This is likely because the global structure, which is not captured by "Total incoming traffic" and "Total traffic" plays a large role in determining the vertices with large stationary distribution values, resulting in large, positive residuals. The plot on the right shows that large residuals between the trajectory feature model is not sufficient for identifying loci since they are interspersed with non-loci residuals.

Table 2 Association between comparison distribution and traffic

Metric	Lin. adj. R^2	RF OOB R^2
Total inc. ~ Total traffic	0.334	0.193
Stat. Dist. ~ Total inc. + Total traffic	0.261	0.230

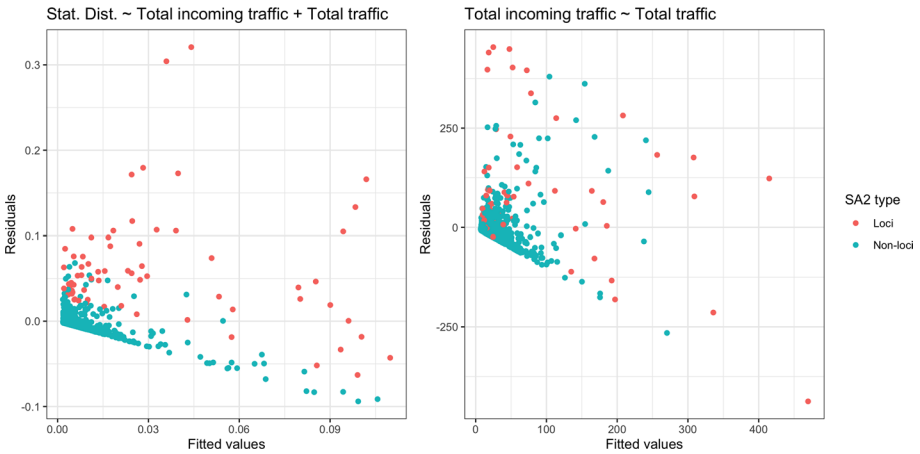


Fig. 16 Residual plots from regressing the stationary distribution on trajectory features (left), and total incoming traffic on total traffic (trajectory features) (right)

Conclusions

As shown, mobility loci identify locations to which people move to and through at a rate higher than expected given the movement and structure of the underlying mobility graph. For the New Zealand commuting data, we identified SA2 areas considered to be mobility loci. These corresponded to specific, high-population and high-movement areas in Auckland and Wellington as well as less populated areas of high mobility. For infrastructure planners, loci with greater total traffic could be seen as places with high "commuting stress" or places where, when they experience delays, have greater impact on commute time on the population than others. Additionally, the two vertex features presented here (stationary distribution and loci status) can be used to study area-level associations with applications in epidemiology and urban planning, among others.

We note that like many analyses conducted on spatially aggregated units like SA2, the results described here are subject to the modifiable areal unit problem where the conclusions can change based on the size and boundaries of the spatial units used. Aggregating data into different-sized units can lead to variations in statistical outcomes, potentially misleading conclusions about spatial relationships and patterns. While the results we arrived at in this analysis make sense contextually, it is possible that had the analysis been done using SA1s (or other spatially aggregated units) as the vertices, the loci identified could be in different regions. Thus, careful consideration must be given to the selection of spatial units to ensure that the findings reflect the underlying spatial scale being considered.

In this analysis, we only had data on home and work locations, which we used to impute the likely commute route to and from work. In the absence of exact time of these commutes, we also assumed that all of the commute happened at the same time (in the morning to work, and in the evening back to home). If data were available at finer spatial or temporal resolutions (e.g. exact route to work and time of day, which could be different on different days), it would be interesting to explore the evolution of the graph over time, and how the dynamics of these loci might vary. Such analyses would also be of interest in other applications such as understanding disease spread. However, given the likely correlated nature of

the location of loci over time, the evolution would need to be modeled to ensure that this characteristic is adequately captured.

As a final note, it should be pointed out that the underlying procedure for identifying loci is by no means limited to mobility graphs and should extend readily to other domains. For example, research in social networks tends to focus on community detection, encoded as an undirected graph with the goal of identifying members of a community. Other areas of application include transportation networks, genomic pathway analysis, or other areas whose representation may be a directed, potentially weighted graph.

Software implementation

All aspects of the analysis presented were implemented using the R Programming Environment (R Core Team 2022). Data formatting and shaping relied on the `dplyr` (Wickham et al. 2022), `tibble` (Müller and Wickham 2022), and `tidyr` (Wickham and Girlich 2022) packages with variable checking performed using the `checkmate` (Lang 2017) package. The construction of routes between home and work SA2's was performed using the `ghroute` (Urbanek 2022) package. Parallel processing, employed to speed-up the permutation procedure and other analyses, used the `foreach` (Microsoft and Weston 2022) package with the multicore backend provided by `doMC` (Revolution Analytics and Weston 2022). Plots were created using `ggplot2` Wickham (2016) and `patchwork` (Pedersen 2020). Spatial visualizations were created using `proj4` (Urbanek 2022), `sf` (Pebesma 2018), and `snippets` (Urbanek 2022) with `RColorBrewer` (Neuwirth 2022) providing color mappings. Mobility graphs were represented and processed and analyzed using a combination of packages `igraph` Csardi and Nepusz (2006), `tidygraph`Pedersen (2022), `Matrix` (Bates et al. 2022), and `graphmobility` (Kane 2022), the last of which encapsulates the novel analysis aspects of this paper. Finally, the `randomForestSRC` (Ishwaran et al. 2008) package was used for the Random Forest analysis.

Author contributions
Kane developed the methods presented, wrote code to perform the presented analyses, and wrote the manuscript. Gilani provided critical feedback and supported methods development. Khusainova provided critical feedback and supported methods development. Urbanek curated the original data, created the commuting directories, and developed the visualizations.

Funding
The first author was supported by the National Science Foundation (NSF) Grant Human Networks and Data Science - Infrastructure (HNDS-I), award numbers 2024335. The second author was supported by the National Science Foundation (NSF) Grant Human Networks and Data Science - Infrastructure (HNDS-I), award numbers 2024233.

Data availibility
No datasets were generated or analysed during the current study.

Declarations

Competing interests
The authors declare no competing interests.

Received: 30 January 2024 Accepted: 8 August 2024
Published online: 06 September 2024

References

Abbasi OR, Alesheikh AA (2018) Exploring the potential of location-based social networks data as proxy variables in collective human mobility prediction models. Arab J Geosci 11:1–14

Alessandretti L (2022) What human mobility data tell us about covid-19 spread. Nat Rev Phys 4(1):12–13
Alessandretti L, Aslak U, Lehmann S (2020) The scales of human mobility. Nature 587(7834):402–407
Barbosa H, Barthelemy M, Ghoshal G, James CR, Lenormand M, Louail T, Menezes R, Ramasco JJ, Simini F, Tomasini M (2018) Human mobility: Models and applications. Phys Rep 734:1–74
Barthélemy M (2011) Spatial networks. Phys Rep 499(1–3):1–101
Bates D, Maechler M, Jagan M (2022) Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.4-1. https://CRAN.R-project.org/package=Matrix
Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc: Ser B (Methodol) 57(1):289–300
Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. Ann Stat 29:1165–1188
Bonaccorsi G, Pierri F, Cinelli M, Flori A, Galeazzi A, Porcelli F, Schmidt AL, Valensise CM, Scala A, Quattrociocchi W et al (2020) Economic and social consequences of human mobility restrictions under covid-19. Proc Natl Acad Sci 117(27):15530–15535
Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste. Studi in onore del professore salvatore ortu carboni, 13–60
Chang J (2007) Stochastic Processes Notes for STAT 251. unpublished book manuscript, accessed 2022-03-25. http://www.stat.yale.edu/~lowpollard/Courses/251.spring2013/Handouts/Chang-notes.pdf
Chen X, Xie J, Xiao C, Lu B, Shan J (2022) Recurrent origin-destination network for exploration of human periodic collective dynamics. Trans GIS 26(1):317–340
Csardi G, Nepusz T (2006) The igraph software package for complex network research. InterJournal Complex Systems, 1695
Dios Ortúzar J, Willumsen LG (2002) Modelling Transport. Wiley, Hoboken
Gilani O, Urbanek S, Kane MJ (2020) Distributions of human exposure to ozone during commuting hours in connecticut using the cellular device network. J Agric Biol Environ Stat 25(1):54–73
Gonzalez MC, Hidalgo CA, Barabasi A-L (2008) Understanding individual human mobility patterns. Nature 453(7196):779–782
GraphHopper: The GraphHopper Directions API Website. https://www.graphhopper.com (2022)
Hommel G (1988) A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75(2):383–386
Hossmann T, Spyropoulos T, Legendre F (2011) A complex network analysis of human mobility. In: 2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 876–881. https://doi.org/10.1109/INFCOMW.2011.5928936
Hyytia E, Lassila P, Virtamo J (2006) Spatial node distribution of the random waypoint mobility model with applications. IEEE Trans Mob Comput 5(6):680–694
Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. Ann Appl Statist 2(3):841–860
Kane MJ, Gilani O (2021) The need to incorporate communities in compartmental models. Stat Interf 14(1):29–32
Kane MJ (2022) Graphmobility: Tools to Analyze Mobility Graphs. R package version 0.1.0. https://github.com/kaneplusplus/graphmobility
Kraemer MU, Yang CH, Gutierrez B, Wu CH, Klein B, Pigott DM, Du Plessis L, Faria NR, Li R, Hanage WP (2020) The effect of human mobility and control measures on the COVID-19 epidemic in China. Science 368(6490):493–7
Lang M (2017) checkmate: Fast argument checks for defensive R programming. R J 9(1):437–445. https://doi.org/10.32614/RJ-2017-028
Microsoft, Weston S (2022) Foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach
Mitsche D, Resta G, Santi P (2014) The random waypoint mobility model with uniform node spatial distribution. Wireless Netw 20(5):1053–1066
Müller K, Wickham H (2022) Tibble: Simple Data Frames. R package version 3.1.7. https://CRAN.R-project.org/package=tibble
Navidi W, Camp T (2004) Stationary distributions for the random waypoint mobility model. IEEE Trans Mob Comput 3(1):99–108
Neuwirth E (2022) RColorBrewer: ColorBrewer Palettes. R package version 1.1-3. https://CRAN.R-project.org/package=RColorBrewer
OpenStreetMap: The OpenStreetMap Website. https://www.openstreetmap.org (2022)
Park YM, Kwan M-P (2017) Individual exposure estimates may be erroneous when spatiotemporal variability of air pollution and human mobility are ignored. Health place 43:85–94
Pebesma E (2018) Simple features for R: standardized support for spatial vector data. R J 10(1):439–446. https://doi.org/10.32614/RJ-2018-009
Pedersen TL (2020) Patchwork: The Composer of Plots. R package version 1.1.1. https://CRAN.R-project.org/package=patchwork
Pedersen TL (2022) Tidygraph: A Tidy API for Graph Manipulation. R package version 1.2.1. https://CRAN.R-project.org/package=tidygraph
R Core Team (2022) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. R Foundation for Statistical Computing. https://www.R-project.org/
Revolution Analytics, Weston, S (2022) doMC: Foreach Parallel Adaptor for 'parallel'. R package version 1.3.8. https://CRAN.R-project.org/package=doMC
Ruan M, Chen X, Zhou H (2019) Centrality prediction based on k-order markov chain in mobile social networks. Peer-to-Peer Netw Appl 12(6):1662–1672
Song C, Qu Z, Blumm N, Barabási A-L (2010) Limits of predictability in human mobility. Science 327(5968):1018–1021
Stats NZ: 2018 Census Main means of travel to work by Statistical Area 2. https://datafinder.stats.govt.nz/table/104720-2018-census-main-means-of-travel-to-work-by-statistical-area-2/ (2020)

Stats NZ: Statistical standard for geographic areas 2018. https://www.stats.govt.nz/methods/statistical-standard-for-geographic-areas-2018 (2017)
Stats NZ: Applying confidentiality rules to 2018 Census data and summary of changes since 2013. https://www.stats.govt.nz/methods/applying-confidentiality-rules-to-2018-census-data-and-summary-of-changes-since-2013 (2019)
Urbanek S (2022) Ghroute: GraphHopper Routing and Navigation. R package version 0.2-0
Urbanek S (2022) Proj4: A Simple Interface to the PROJ.4 Cartographic Projections Library. R package version 1.0-11. http://www.rforge.net/proj4/
Urbanek S (2022) Snippets: Code Snippets, Mostly Visualization-related. R package version 0.1-2. http://www.rforge.net/snippets/
Wickham H (2016) Ggplot2: elegant graphics for data analysis. Springer, Cham
Wickham H, François R, Henry L, Müller K (2022) Dplyr: A Grammar of Data Manipulation. R package version 1.0.9. https://CRAN.R-project.org/package=dplyr
Wickham H, Girlich M (2022) Tidyr: Tidy Messy Data. R package version 1.2.0. https://CRAN.R-project.org/package=tidyr
Yuan Y, Raubal M (2016) Analyzing the distribution of human activity space from mobile phone usage: an individual and urban-oriented study. Int J Geogr Inf Sci 30(8):1594–1621. https://doi.org/10.1080/13658816.2016.1143555

Publisher's Note