What lies beneath: a spatial mosaic of Zika virus transmission in the 2015-2016 epidemic in Colombia

T. Alex Perkins[1,*,$], Isabel Rodriguez-Barraquer[2,*], Carrie Manore[3,*], Amir S. Siraj[1], Guido España[1], Christopher M. Barker[4], Michael A. Johansson[5,6], Robert C. Reiner[7,*,$]

[1] Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, Notre Dame, IN, USA, taperkins@nd.edu, asiraj@nd.edu, guido.espana@nd.edu
[2] Department of Medicine, University of California, San Francisco, CA, USA, Isabel.Rodriguez@ucsf.edu
[3] Theoretical Biology and Biophysics, Los Alamos National Laboratory, Los Alamos, NM, USA, cbearmath@gmail.com
[4] Department of Pathology, Microbiology, and Immunology, University of California, Davis, CA, USA, cmbarker@ucdavis.edu
[5] Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, San Juan, PR, USA, mjohansson@cdc.gov
[6] Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health, Boston, MA, USA
[7] Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA, bcreiner@uw.edu

* Contributed equally
[$] Corresponding authors

ABSTRACT

Background: Temporal incidence patterns provide a crucial window into the dynamics of emerging infectious diseases, yet their utility is limited by the spatially aggregated form in which they are often presented. Weekly incidence data from the 2015-2016 Zika epidemic were available only at the national level for most countries in the Americas. One exception was Colombia, where data at departmental and municipal scales were made publicly available in real time, providing an opportunity to assess the degree to which national-level data are reflective of temporal patterns at local levels.

Methods: To characterize differences in epidemic trajectories, our analysis centered on classifying proportional cumulative incidence curves according to six features at three levels of spatial aggregation. This analysis used the partitioning around medoids algorithm to assign departments and municipalities to groups based on these six characteristics. Examination of the features that differentiated these groups and exploration of their temporal and spatial patterns were performed. Simulations from a stochastic transmission model provided data that were used to assess the extent to which groups identified by the classification algorithm could be associated with differences in underlying drivers of transmission.

Results: The timing of departmental-level epidemic peaks varied by three months, and departmental-level estimates of the time-varying reproduction number, $R(t)$, showed patterns that were distinct from a national-level estimate. The classification algorithm identified moderate support for two to three clusters at the departmental level and somewhat stronger support for this at the municipal level. Variability in epidemic duration, the length of the tail of the epidemic, and the consistency of cumulative incidence data with a cumulative normal distribution function made the greatest contributions to distinctions across these groups. Applying the classification algorithm to simulated data showed that municipalities with basic reproduction number, R_0, greater than 1 were consistently associated with a particular group. Municipalities with $R_0 < 1$ displayed more diverse patterns, although in this case that may be due to simplifications of how the model represented spatial interaction among municipalities.

Conclusions: The diversity of temporal incidence patterns at local scales uncovered by this analysis underscores the value of spatially disaggregated data and the importance of locally tailored strategies for responding to emerging infectious diseases.

Keywords: Colombia; data analytics; emerging infectious disease; epidemic; forecasting; mathematical modeling; spatial analysis; surveillance data; time series; Zika

BACKGROUND

Time series have been used for many years to make inferences about processes that shape the dynamics of a wide range of systems [1]. This long history has resulted in appreciation of a number of common challenges for time series analysis [2]. One such challenge is disentangling the effects of multiple interacting forces, which can include both extrinsic forces, such as weather, and intrinsic forces, such as immune-mediated feedbacks [3,4]. An even more fundamental challenge lies in defining the time series in the first place, especially with respect to space [5]. The question is, at what spatial scale should epidemiological data be aggregated for time series analysis?

In practice, the spatial scale at which data are aggregated to form a time series is more often dictated by the scale at which data are available than by the scale that is optimal for inference or prediction. For example, during the recent invasions of chikungunya virus (CHIKV) and then Zika virus (ZIKV) across the Americas, the Pan American Health Organization published weekly case reports aggregated nationally. Despite an abundance of evidence that chikungunya and dengue viruses – another virus transmitted by *Aedes aegypti* mosquitoes – are characterized by spatially focal transmission [6,7], applications ranging from estimation of time-varying reproduction numbers [8] to forecasting [9] have utilized data aggregated at national scales for countries as vast and spatially heterogeneous as Brazil and Mexico.

One exception to the public availability of data only at coarse, highly aggregated scales is Colombia, where routine surveillance of Zika was reported on a weekly basis for each of the country's 1,123 municipalities during the 2015-2016 epidemic [10]. Although these case reports

are underestimates of the true extent of transmission of many infectious diseases, particularly those with high proportions of asymptomatic infections, they still provide a uniquely valuable resource given the paucity of publicly available data at similar scales in other countries [11]. Such data are particularly valuable for Zika, given that a range of spatial scales are relevant for activities related to its prevention and control. On the one hand, vector control activities are planned and budgeted on multiple administrative levels but must be targeted on a very local level. On the other hand, communications, surveillance, and possible vaccination programs are generally planned and implemented only on larger administrative scales.

Our goal in this study was to utilize this unique data set on the ZIKV invasion of Colombia to perform a case study on the characteristics of temporal incidence patterns at different spatial scales in the context of an emerging infectious disease. To do so, we took a three-part approach. First, we performed a descriptive analysis of time series of weekly case reports at three distinct scales in Colombia: national, departmental, and municipal. Second, we performed a classification analysis of proportional cumulative incidence curves at departmental and municipal scales to identify distinct patterns of temporal dynamics at each of these scales. Third, we repeated the classification analysis for data simulated with a mechanistic model for ZIKV transmission in Colombia to determine the extent to which distinct statistical patterns in temporal incidence may reflect distinct driving processes.

METHODS

Data

The weekly number of Zika cases, by municipality, was reconstructed using two data sources. The main data source was a website [12] of the Colombian National Institute of Health (Instituto Nacional de Salud, INS) where the official weekly reports on the cumulative number of suspected and confirmed Zika cases for each municipality were published beginning in early 2016. While the peak of the Colombian epidemic occurred in 2016, a significant number of cases were reported during 2015. To capture this initial portion of the epidemic, we used an additional data source, also available on the INS website. Unfortunately, the number of cases reported in the latter data source seemed to consistently underreport the total number of cases reported by the INS at the national scale. For example, while the official data source reports a cumulative number of 11,712 cases by the end of 2015, this secondary source only reports 3,875 cases for this same period. Therefore, to reconstruct the 2015 portion of the epidemic while accounting for the better known total number of cases, we multiplied the weekly 2015 data by a correction factor. This correction factor was calculated as the ratio between the cumulative number of cases reported by each municipality up to the first week of 2016 according to the official source and the alternative source.

For the simulation model, we used two data sources. First, we obtained gridded population data across Colombia for 2015 at a resolution of 3 arc seconds (~93 m) from WorldPop [13]. We summed these raster data at the municipal level as defined by GIS shapefiles from the National Geographical Information System of Colombia [14]. Second, we based estimates of R_0 on a set

of ZIKV epidemic size projections for Latin America made early in the epidemic using relationships between environmental variables and transmission metrics [15]. To obtain a single value of R_0 for each municipality, we took a weighted sum of the R_0 raster at 5 km x 5 km resolution weighted by a population raster aggregated to that scale. We calibrated these R_0 estimates to observed dynamics in Colombia by scaling municipal values of R_0 from [15] by a constant (2.72) such that the value for Girardot matched an estimate of 4.61 derived from an analysis of temporal incidence patterns there [16].

Descriptive analysis of weekly case reports

We performed two preliminary analyses of differences in weekly case report patterns at different scales of spatial aggregation. First, we generated a barplot of national case reports color-coded by which of 32 departments those national cases arose from. Likewise, for each of those departments, we generated a barplot of departmental case reports color-coded by which of its municipalities those departmental cases arose from. Second, we made estimates of the time-varying effective reproduction number, $R(t)$, for each time series. Following Ferguson et al. [8], we used the EstimateR function from the EpiEstim library [17] in R to estimate $R(t)$ for each time series based on the method introduced by Cori et al. [18].

Classification analysis of cumulative incidence curves

We focused our analysis on cumulative, rather than raw, incidence because of the extreme variability in raw incidence patterns in this data set. With raw incidence, time series with a small number of cases appear extremely jagged, and temporal patterns can be difficult to extract. With proportional cumulative incidence, vastly different temporal patterns are more readily comparable, because they all begin at 0 and end at 1 but arrive there by different paths. Others [19] have criticized the use of cumulative incidence data from epidemics, although these criticisms mostly pertain to parameter estimation and forecasting, neither of which we do here. Rather, our goal was to perform a descriptive analysis of diversity in the temporal patterns of an epidemic as viewed from many different perspectives spatially.

The cumulative incidence curves that we examined were proportional such that they all reached 1 at the time the last case was reported in a given area. Mathematically, for weekly reported Zika incidence $I_{i,t}$ in location i in week t, we calculated proportional cumulative incidence as

$$C_{i,t} = \frac{\sum_{\tau \le t} I_{i,\tau}}{\sum_{\tau} I_{i,\tau}}. \qquad (1)$$

We excluded areas from our analysis that reported no Zika cases.

As a basis for classifying cumulative incidence curves, we defined six features of these curves that we hypothesized represent dimensions in which curves from different areas vary (Table 1). We defined four of these features in reference to cumulative normal density curves, $\hat{C}_i(t)$, that we fitted to each $C_{i,t}$. This involved estimating mean and standard deviation parameters of $\hat{C}_i(t)$ for each $C_{i,t}$ on the basis of least squares using the optim function in R. We chose these

features because they provided a way to quantify the duration of local epidemics (small F_{SD}, short $F_{\Delta t}$ = short epidemic), to capture whether epidemics appeared strongly locally driven (low F_{R^2}, large F_0= sporadic transmission fueled by importation), and to characterize shapes that deviated substantially from those predicted by simple epidemic models ($F_{5\%}$ and $F_{95\%}$ near zero = "SIR-like" epidemic). Although these idealized scenarios motivated the selection of these features, the fact that all six features were calculated for each $C_{i,t}$ meant that we were able to capture a wide range of patterns in between these extremes.

Symbol	Definition
F_{SD}	Standard deviation of $\hat{C}_i(t)$
F_{R^2}	R^2 between $C_{i,t}$ and $\hat{C}_i(t)$
$F_{5\%}$	Difference between the 5% quantile of $C_{i,t}$ and the 5% quantile of $\hat{C}_i(t)$
$F_{95\%}$	Difference between the 95% quantile of $C_{i,t}$ and the 95% quantile of $\hat{C}_i(t)$
$F_{\Delta t}$	Weeks between first and last non-zero $C_{i,t}$
F_0	Weeks with $C_{i,t}$ = 0 between first and last non-zero $C_{i,t}$

Table 1. Features of proportional cumulative incidence curves used for classification analysis.

We explored variation in $C_{i,t}$ at both departmental and municipal scales. To describe how variation in $C_{i,t}$ curves at these scales was distributed across the six-dimensional feature space, we performed a partitioning around medoids (PAM) clustering analysis [20] on scaled values of the features using the pam function in the cluster library [21] in R. This algorithm identifies medoids of *k* groups that minimize the sum of distances between each medoid and all group members. We performed this analysis for values of *k* ranging 2-10 and compared groupings for different values of *k* on the basis of their average silhouette values. A silhouette value describes how much more dissimilar a point is from points in the next most similar group compared to points in its own group [22]. An ideal classification would be indicated by silhouette values for data points in all groupings close to 1. Silhouette values nearer to or below 0 indicate that points do not cluster particularly well with the group to which they are assigned.

Simulation of epidemic curves to elucidate driving processes

To aid in the interpretation of the classification analysis of observed patterns of temporal incidence, we performed an identical analysis of simulated patterns of temporal incidence. The value of doing so is that it provides an opportunity to determine the extent to which groups identified by the classification analysis might reflect meaningful differences among those groups in terms of transmission processes and their drivers. In other words, we viewed this exercise as a form of validation of the overall approach of performing a classification analysis on the features of proportional cumulative incidence patterns listed in Table 1.

We simulated a data set comparable to the observed data using an R implementation of the ZIKV transmission model described by Ferguson et al. [8] parameterized to match the municipal-level R_0 values described at the end of the previous section. The model by Ferguson et al. had a number of attractive features, including plausible values of a number of parameters common to ZIKV transmission models, realistic accounting of the timing of transmission-relevant processes in mosquitoes and humans, seasonal variation in transmission, and the ability to capture multiple forms of stochasticity associated with transmission and surveillance. In brief, the model assumes that humans transition from a susceptible compartment into a recovered and immune compartment following a period of incubation and infectiousness and that mosquitoes become infectious and remain so following bites of infectious humans and a seasonally variable incubation period. Mosquito population density is also seasonally variable, driven by seasonal variation in larval carrying capacity and adult mortality. A full description of the model can be found in the paper by Ferguson et al. [8].

To apply this model to Colombia, we used municipal-level human population sizes derived from WorldPop [13] and adjusted seasonally averaged mosquito densities such that seasonally averaged values of R_0 matched our municipal-level R_0 estimates. Another departure from the original model that we made was to remove explicit spatial coupling, given the complexity of doing so realistically for all 1,122 municipalities in Colombia. Instead, we simulated imported infections (i.e., infections acquired outside a given municipality) to occur at a daily per capita rate that was proportional to a normal probability density function fitted to the temporal pattern of national-scale incidence (timing of national-scale incidence: mean = 32.57 weeks after the first reported case, standard deviation = 8.85 weeks). This time-varying importation function was scaled by a value of 1.55×10^{-3}, which along with an assumed reporting rate of 11.5% [23] allowed us to approximately match the national total of 85,353 suspected Zika cases. Also, given that our interest was in short-term dynamics rather than long-term dynamics as in [8], we removed human age stratification from the model.

RESULTS

Descriptive analysis of weekly case reports

As a whole, the temporal incidence pattern at the national level was consistent with a typical epidemic trajectory, marked by an increase over approximately five months, a peak around the beginning of February 2016, and a steady decline thereafter over a period of approximately eight months (Fig. 1A). Under a standard set of assumptions about epidemic dynamics, this incidence pattern can be used to estimate the temporal trajectory of the effective reproduction number, $R(t)$ [18]. Applying this technique at the national level yielded estimates of $R(t)$ that began high (range: 1.5-3.5 for the first four months) and gradually declined below 1 by the time the epidemic concluded (Fig. 1A), all of which is consistent with standard expectations for an epidemic of an immunizing pathogen in an immunologically naive host population.

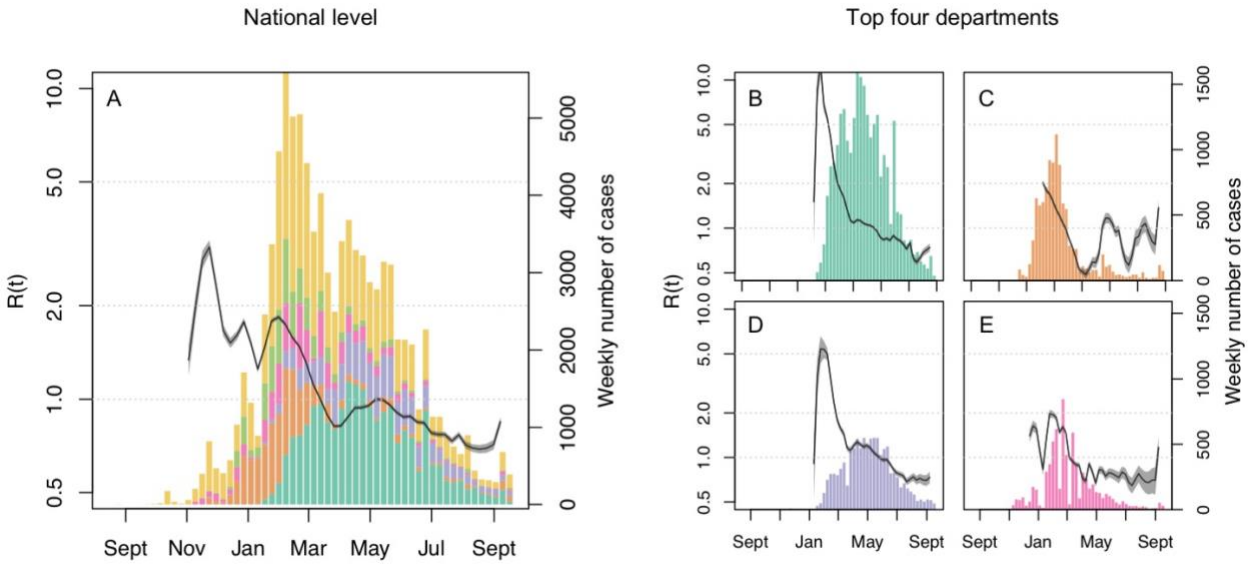

Figure 1. Weekly Zika case reports at the national level (A) and for each of the four departments with the largest case report totals (B: Valle del Cauca; C: Norte de Santander; D: Santander; E: Tolima). Colors match across A and B-E, with the addition of yellow in A that includes all departments other than those in B-E. Time-varying estimates of the effective reproduction number, $R(t)$, are shown in each panel.

Examination of temporal incidence patterns for each of the four largest departments in terms of total incidence (Valle del Cauca, Norte de Santander, Santander, Tolima) showed that patterns at the departmental level were quite different than those at the national level. First, the timing of peak incidence in the departments in Fig. 1B-1E varied by around three months. Second, the shapes of the incidence patterns in these departments varied, with Valle del Cauca and Santander (Fig. 1B & 1D) showing high incidence sustained over a period of several months and Norte de Santander and Tolima (Fig. 1C & 1E) showing sharper peaks trailed by relatively low incidence for several months after.

This high degree of variability in temporal incidence patterns had substantial impacts on estimates of $R(t)$. At the national level, $R(t)$ estimates never exceeded 3.5, whereas in Santander $R(t)$ was estimated to exceed 5 (Fig. 1D) and in Valle del Cauca it was estimated to exceed 10 (Fig. 1B), due in both cases to more rapid increases in incidence at the departmental level than the national level. In Norte de Santander, $R(t)$ appeared to twice fall well below 1 but then quickly rise back above this critical threshold value (Fig. 1C).

Examination of temporal incidence patterns at the municipal scale revealed even more variability in temporal incidence patterns than at the department level. In the department of Norte de Santander (Fig. 1C), for example, it was clear that one municipality dominated the departmental pattern (Fig. 2A). The municipalities with the second and third highest incidence both experienced short, unimodal patterns of incidence during the first two months, but incidence patterns thereafter were mostly low and erratic (Fig. 2B & 2C). Other municipalities in

the department had only low, erratic incidence with no sign of a distinct epidemic (e.g., Fig. 1D). With the exception of the first few weeks of transmission, estimates of $R(t)$ at the municipal level were characterized by erratic fluctuations and much larger uncertainty than was apparent at the departmental or national level (Fig. 2).

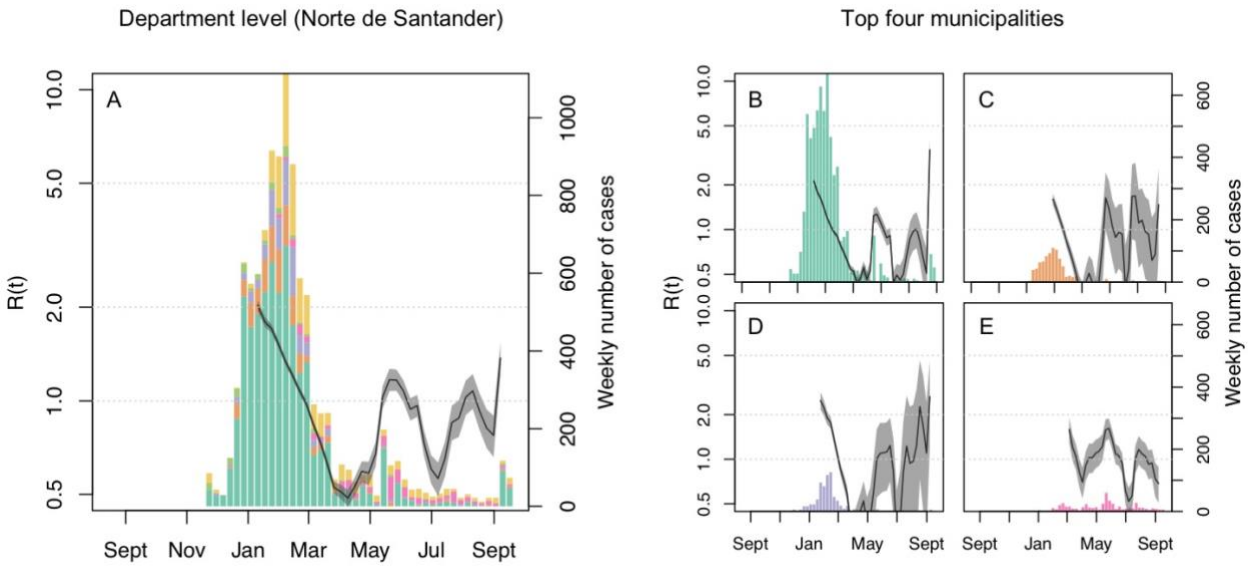

Figure 2. Weekly Zika case reports at the departmental level for Norte de Santander (A) and for each of the four municipalities with the largest case report totals (B: Cucuta; C: Villa del Rosario; D: Los Patios; E: Ocaña). Colors match across A and B-E, with the addition of yellow in A that includes all municipalities other than those in B-E. Time-varying estimates of the effective reproduction number, $R(t)$, are shown in each panel.

Classification analysis of cumulative incidence curves

At the departmental level, there was only modest clustering overall, with the highest average silhouette value corresponding to two groups (0.256), a slightly lower value for three groups (0.254), and falling no lower than 0.201 for up to ten groups (Fig. S1). F_{SD} and $F_{95\%}$ were the features that were most important for distinguishing two groups (Fig. S2), and $F_{\Delta t}$ contributed further to distinguishing three groups (Fig. S3). Differences in F_{SD} were associated with a difference of approximately two months in the time elapsed between the attainment of 5% and 80% of cumulative incidence (Fig. 3, left: blue longer than red), and differences in $F_{95\%}$ were associated with a difference of approximately two months in the time elapsed between the attainment of 80% and 99% of cumulative incidence, but for different groups (Fig. 3, left: red longer than blue). Overall, this meant that the time elapsed between attainment of 5% and 99% of cumulative incidence for both groups was similar, but with one group experiencing epidemics that were fast initially but slow to finish and another group experiencing epidemics that were slower initially but finished more quickly. These patterns were clearest for the curves associated with the medoid of each group (Fig. 3) but were generally apparent for the curves associated with the groups as a whole (Fig. 4). Spatially, groups tended to cluster along northern, central,

and southern strata (Fig. 5, left), with incidence-weighted cartographs showing that the epidemic was mostly dominated by distinct northern and central strata (Fig. 5, top right).

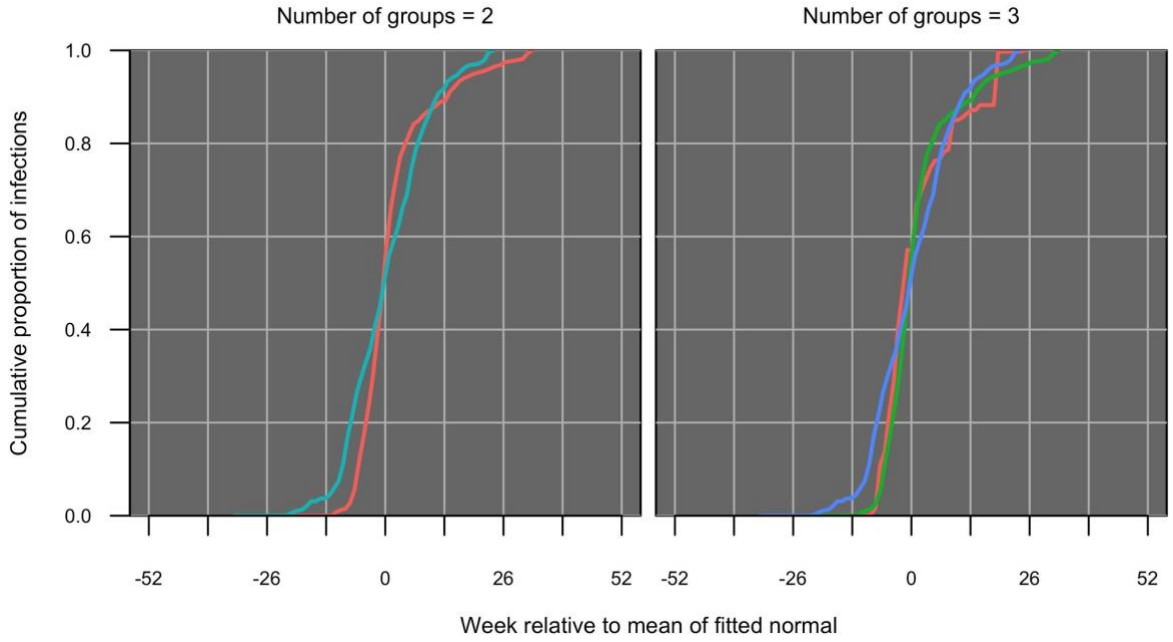

Figure 3. Proportional cumulative incidence curves at the departmental level with two (left) or three (right) groups. Only one representative curve is shown for each group, with that curve being chosen on the basis of being associated with the medoid of its group.

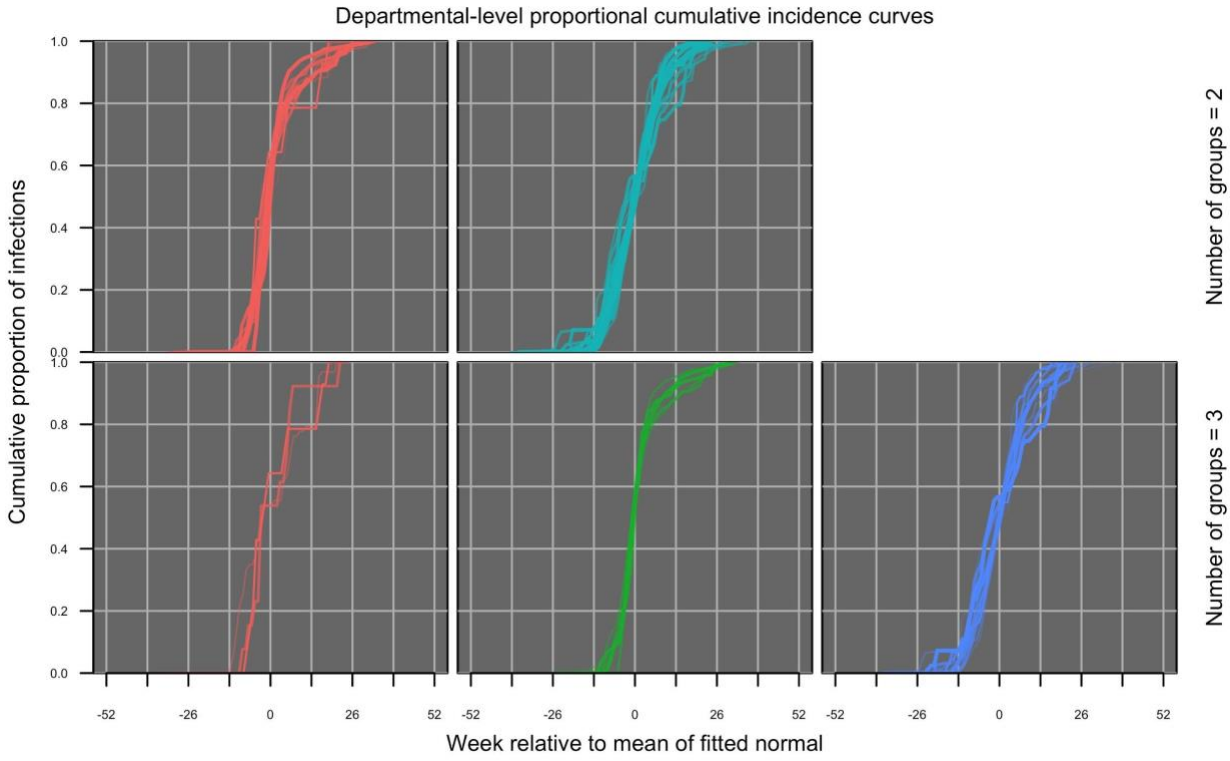

Figure 4. Proportional cumulative incidence curves at the departmental level with two (top) or three (bottom) groups. Within each row, groups are distinguished by color.

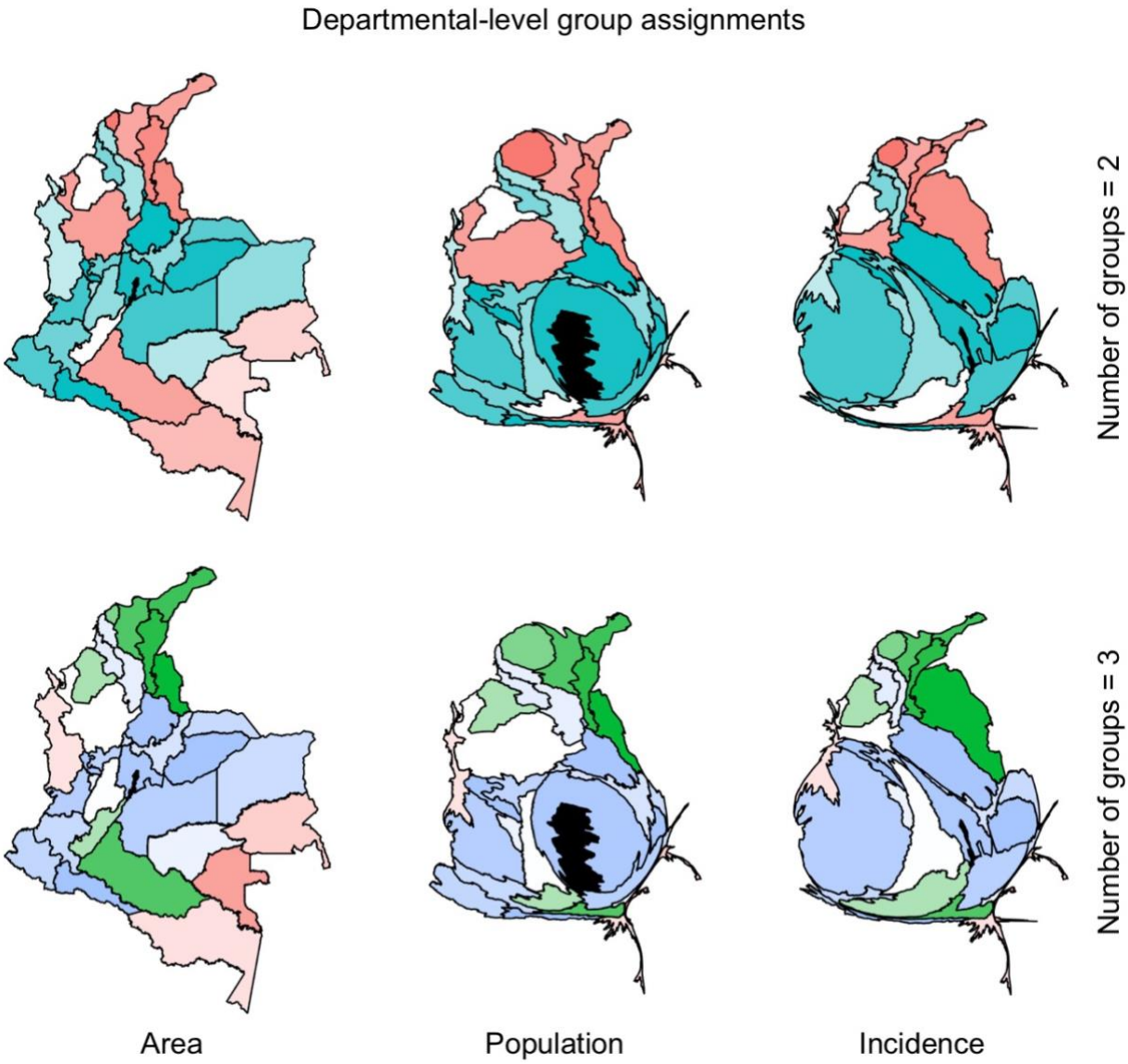

Figure 5. Cartograms at the departmental level weighted by area (left), population (center), and incidence (right). Department assignments to two (top) and three (bottom) groups are indicated by color, with transparency inversely proportional to silhouette value. The one department (Bogotá) with zero incidence is indicated in black and given a weight equivalent to 1/5 of a case to allow for its inclusion in the right column.

There was somewhat stronger clustering at the municipal level, with the highest average silhouette value corresponding to three groups (0.352), somewhat lower values for five and six groups (0.334, 0.326), and no lower than 0.297 for up to ten groups (Fig. S4). $F_{\Delta t}$ and F_{SD} were the features that were most important in distinguishing two groups (Fig. S5), $F_{95\%}$ made additional contributions to distinguishing three groups (Fig. S6), and F_{R^2} contributed to

distinguishing four groups (Fig. S7). Proportional cumulative incidence curves for the group with short $F_{\Delta t}$ and small F_{SD} were the most visually distinct group and remained relatively consistent regardless of the number of groups (Fig. 6). Some differences among the other groups were also apparent in the proportional cumulative incidence curves, with some having a long tail (Fig. 6, middle: green) or two discrete jumps (Fig. 6, middle: blue). The timing of discrete jumps varied across municipalities, but curves within a group otherwise resembled the curve associated with the medoid for that group (Fig. 7). Spatially, departments generally consisted of a mixture of municipalities from different groups, and the prominence of some groups in the cartograms varied depending on whether the cartograms were weighted by area, population, or incidence (Fig. 8). The cartograms weighted by population showed that a sizeable portion of the population lives in cities that had no reported cases, such as Medellín and Bogotá (Fig. 8, black in the center column). Among municipalities that did have reported cases, the cartograms weighted by incidence showed that a relatively large proportion of reported cases came from municipal-level epidemics characterized by large $F_{\Delta t}$ and F_{SD} (Fig. 8, right column).

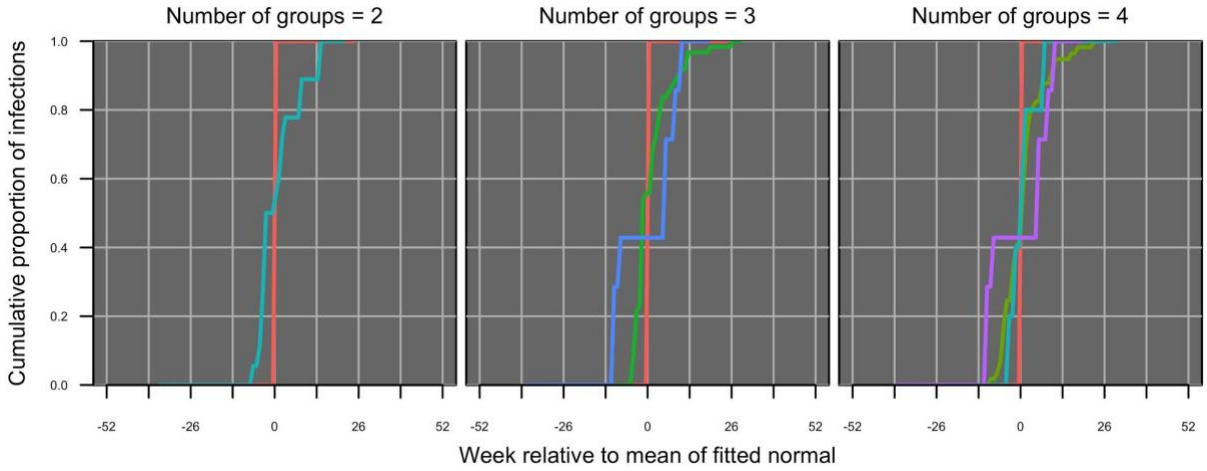

Figure 6. Proportional cumulative incidence curves at the municipal level with two (left), three (middle), or four (right) groups. Only one representative curve is shown for each group, with that curve being chosen on the basis of being associated with the medoid of its group.

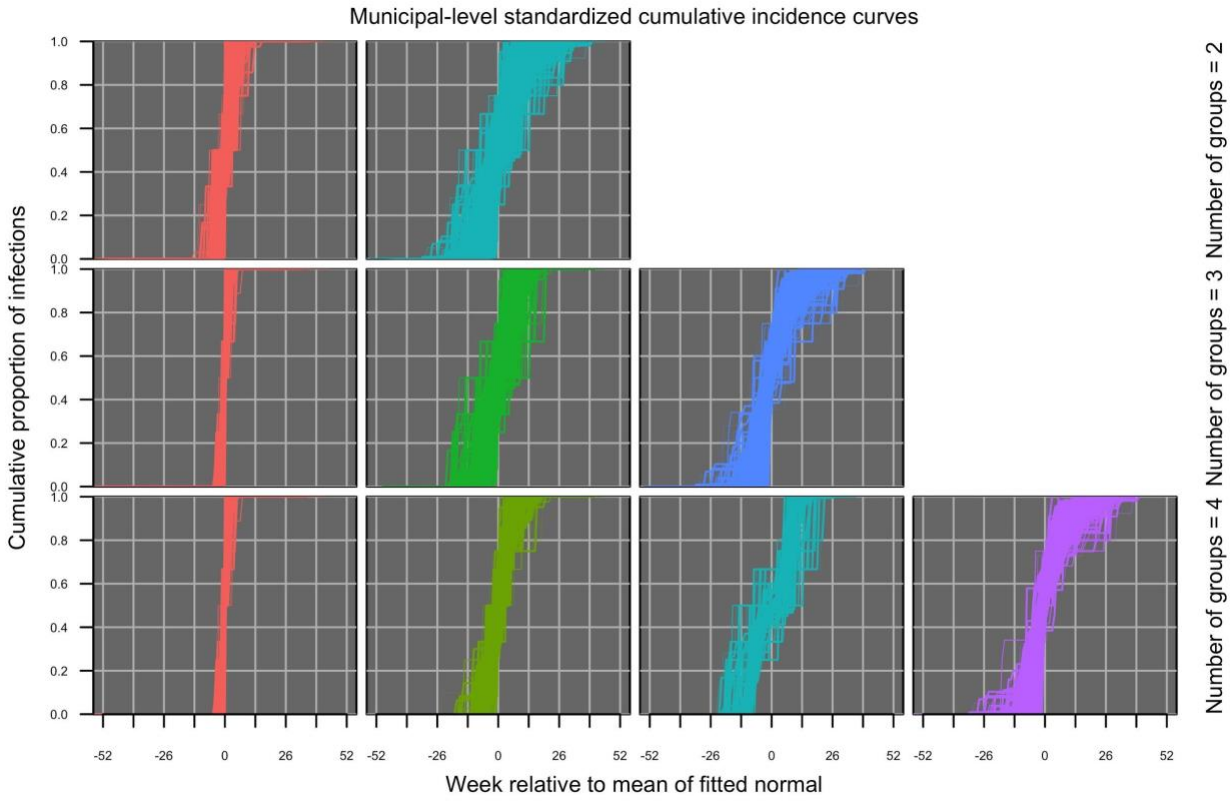

Figure 7. proportional cumulative incidence curves at the municipal level with two (top), three (middle), or four (bottom) groups. Within each row, groups are distinguished by color.

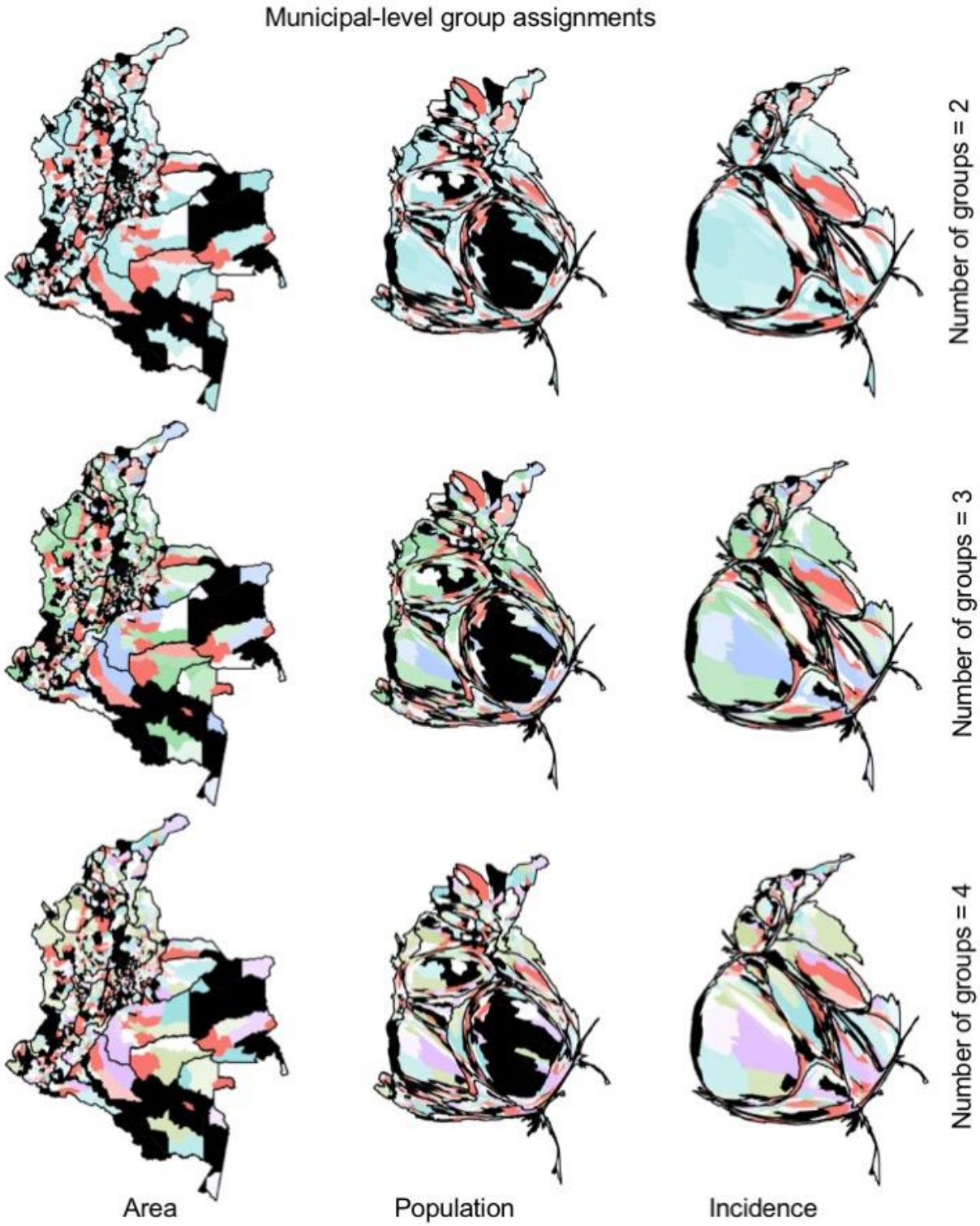

Figure 8. Cartograms at the municipal level weighted by area (left), population (center), and incidence (right). Municipality assignments to two (top), three (middle), and four (bottom) groups are indicated by color, with transparency inversely proportional to silhouette value. Municipalities with zero incidence are indicated in black and were given a weight equivalent to 1/5 of a case to allow for their inclusion in the right column.

Simulation of epidemic curves to elucidate driving processes

We focused our analysis of simulated data at the municipal level given that the simulation model was not equipped to simulate transmission between municipalities, which is likely important for recreating departmental-level patterns. Overall, our model parameterization assumed that $R_0 > 1$ in 34.6% of municipalities. A total of 12.6% (range: 10.4-14.1%) of municipalities had zero simulated cases, with 99.0% (range: 97.0-100.0%) of those having $R_0 < 1$.

Out of 100 simulated datasets, the classification algorithm selected two groups eight times, three groups 80 times, and five and six groups four times each. Average silhouette value was 0.313 (range: 0.288-0.347) when there were two groups and 0.327 (range: 0.291-0.352) when there were three groups (see Fig. S8 for a representative silhouette plot from a randomly selected simulated dataset). Although this indicates a modest preference of the algorithm for three groups, we focused subsequent analyses on the two-group classification due to our desire to evaluate the correspondence between groups selected by the classification analysis and groups defined by R_0 above or below 1.

With the two-group classification, 99.1% (range: 90.3-100.0%) of municipalities with $R_0 > 1$ were placed into the group characterized by larger $F_{\Delta t}$ and F_{SD}. Of the municipalities with $R_0 < 1$, 74.0% (range: 36.3-80.5%) were also placed into that group, with the others placed into the group with smaller $F_{\Delta t}$ and F_{SD} (see Fig. S9 for an example from a randomly selected simulated dataset). When municipalities were classified into three groups, a new group characterized by moderately low $F_{\Delta t}$ and F_{SD} and negative $F_{95\%}$ contained 18.8% (range: 0.2-36.1%) of municipalities with $R_0 > 1$ and 44.7% (range: 23.0-56.5%) with $R_0 < 1$ (see Fig. S10 for an example from a randomly selected simulated dataset). In the presence of this third group, 79.9% (range: 63.4-89.7%) of municipalities with $R_0 > 1$ and 32.1% (range: 22.8-38.8%) with $R_0 < 1$ were placed into the group characterized by larger $F_{\Delta t}$ and F_{SD}.

Visual inspection of five simulated datasets showed that the proportional cumulative incidence curves of municipalities placed in the group characterized by large $F_{\Delta t}$ and F_{SD} generally resembled the curves of municipalities with $R_0 > 1$ (Fig. 9, red). In contrast, proportional cumulative incidence curves of municipalities with $R_0 < 1$ were more diverse than those placed in the group characterized by low $F_{\Delta t}$ and F_{SD} (Fig. 9, blue). A similar pattern was apparent spatially, with municipalities placed in the group characterized by large $F_{\Delta t}$ and F_{SD} generally overlapping with municipalities with $R_0 > 1$, but municipalities with $R_0 < 1$ frequently placed in the group characterized by large $F_{\Delta t}$ and F_{SD} (Fig. 10).

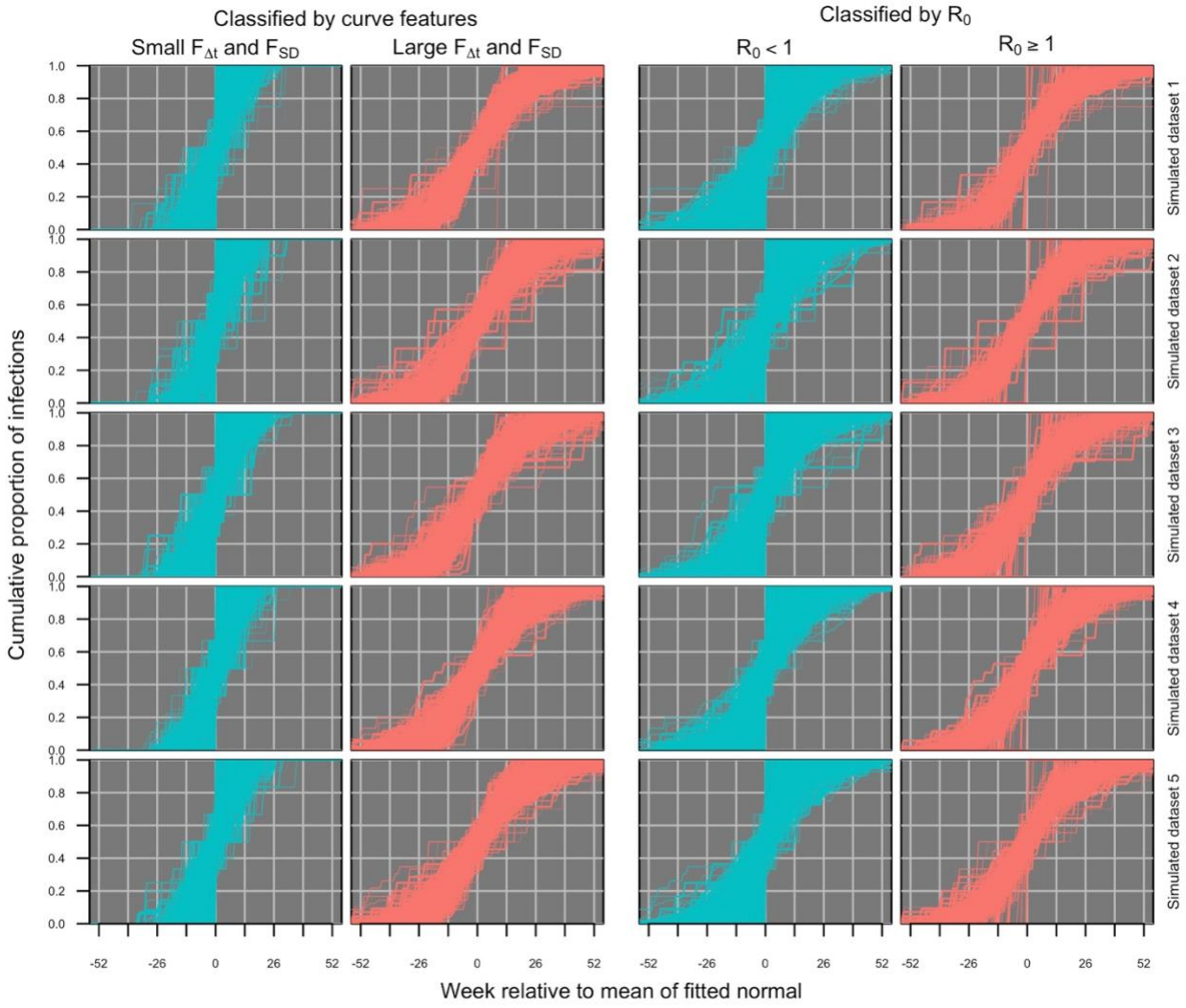

Figure 9. proportional cumulative incidence curves at the municipal level from five randomly selected simulated datasets. The left two columns show two different groups classified by the curve classification algorithm, and the right two columns show two different groups defined by whether those municipalities have a R_0 above or below 1.

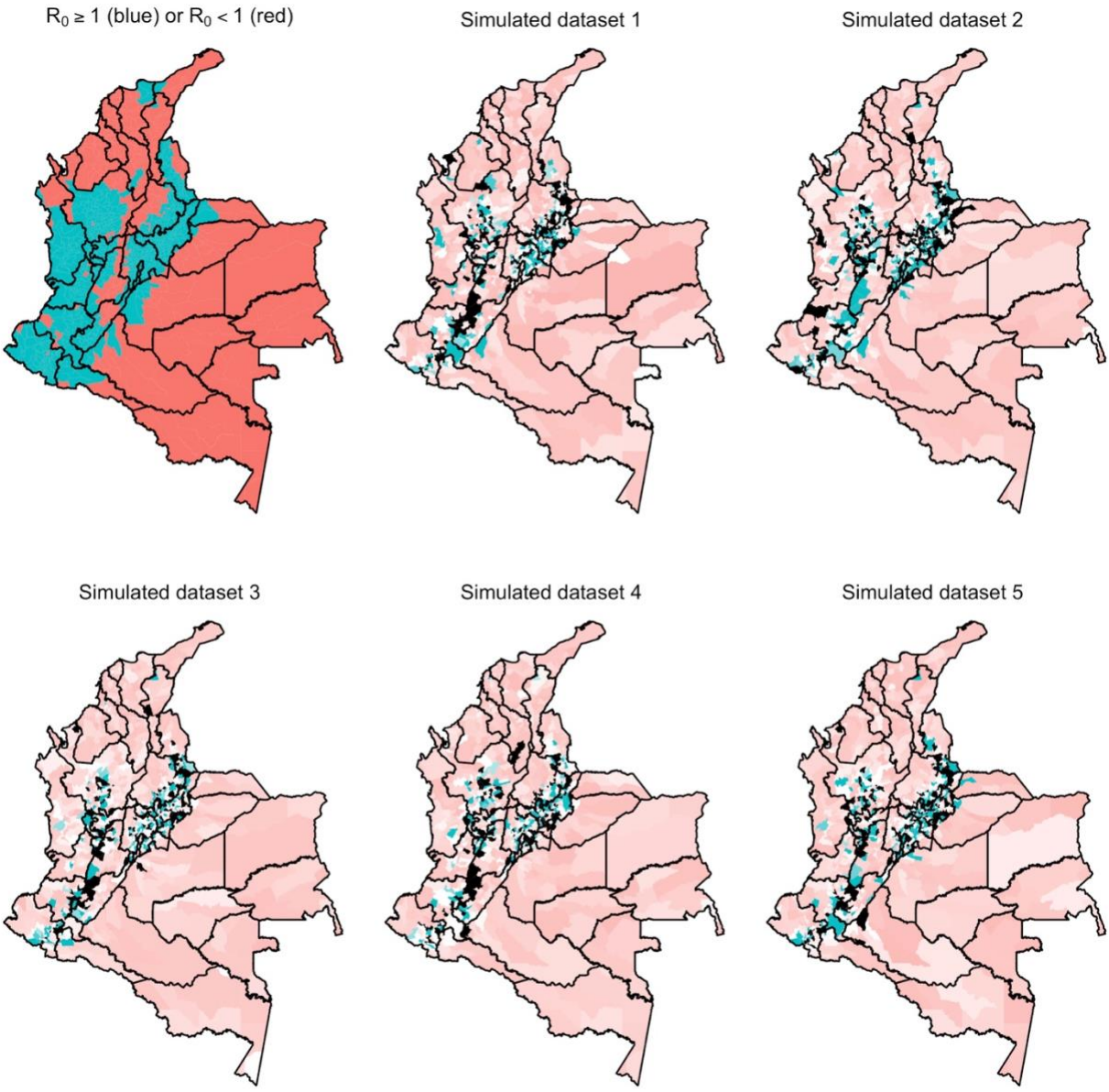

Figure 10. Cartograms at the municipal level weighted by area based on model simulations. Each municipality's status as having $R_0 > 1$ (red) or $R_0 < 1$ (blue) is indicated in the top left panel. In each of five simulated datasets shown in the other panels, municipality assignments to two groups are indicated by color, with transparency inversely proportional to silhouette value. Black indicates that no reported cases were simulated for that municipality.

DISCUSSION

Temporal incidence patterns play a vital role in inferring pathogen transmission dynamics and drivers thereof. By analyzing data from the 2015-2016 Zika epidemic in Colombia, we showed that these patterns can appear very different depending on the spatial scale at which incidence data are aggregated. Whereas national-level patterns appeared to follow a unimodal pattern

consistent with behavior of standard epidemic models, departmental-level patterns were somewhat more varied and municipal-level patterns were the most varied. Combining these observations with a formal classification of temporal incidence patterns and a model-based exploration of mechanisms capable of generating those patterns, we deduced that there is distinct variation in temporal incidence patterns subnationally and that much of that variation may be driven by spatial variation in local transmission potential. Spatial contiguity of areas classified into the same groups was observable to some extent at the departmental level but was generally not observed at the municipal level, suggesting that there are underlying spatial drivers of the variation that we observed in temporal incidence patterns.

Similar to our findings of differing dynamics at municipal and departmental scales, theoretical analyses of a range of ecological models have proposed that dynamics approach deterministic behavior as spatial scales grow larger and data become increasingly more aggregated [24]. Methods based on long-term dynamics have been proposed for identifying the scales at which behavior transitions from stochastic to deterministic in models of plant competition and predator-prey interactions [25,26]. Epidemics, however, are inherently transient in nature [2], leaving open the question of how best to define characteristic spatial scales in that context. It is certainly the case that the data from Colombia that we examined displayed greater stochasticity at finer spatial scales. At the same time, the greater variability in temporal patterns that we observed at finer scales suggests that models that aspire to a deterministic representation of behavior at coarser scales must account for spatial structure at finer scales. Indeed, a recent attempt to fit a national-scale transmission model to national-scale time series of Zika case reports from Colombia showed that ignoring subnational spatial structure inhibited that model's fit to the data [27]. A theoretical exploration of similar issues concluded that the scale at which spatial structure must be modeled explicitly is expected to vary by pathogen and geographic context, with less mobile pathogens requiring explicit spatial representation at finer scales [28].

Both stochasticity and spatial interaction are expected to contribute to variability in temporal dynamics at local scales [29]. For some municipalities, temporal incidence patterns appeared to be dominated by stochasticity (e.g., those with discrete jumps). For others, there were implications for a role of spatial interaction (e.g., those with two sharp increases or a long tail). Whereas our simulation model was realistic with respect to demography and the inclusion of spatiotemporal variability in local transmission, it made the very simplistic assumption that importation patterns have identical timing and magnitude in all municipalities. This may have caused municipalities with $R_0 < 1$, particularly those with larger populations, to display incidence patterns that simply reflected the national trend used to drive importation. Analyses of subnational spatiotemporal dynamics in a range of contexts show that importation patterns vary substantially over time and as a function of regional connectivity or being positioned on an international border [30–33]. Future work that includes more realistic spatial interaction among subnational units could be helpful for resolving the hypothesis proposed here about the importance of spatial interaction in shaping temporal incidence patterns at each of the spatial scales that we considered.

Our finding of differences in temporal incidence patterns at different spatial scales raises questions about infectious disease forecasting that deserve further consideration, particularly at national and other highly aggregated scales. Analyses of alternative models used to forecast Ebola virus disease incidence in the West African epidemic showed that the accuracy of forecasts can be extremely sensitive to model misspecification [19]. More recent work addressing the predictive limits of forecasting for a number of infectious diseases has suggested that appropriate model structure may itself change over the course of an epidemic, with overly rigid model structures being constrained in their ability to accurately forecast future incidence patterns [34]. Our findings –specifically, that temporal incidence patterns vary across municipalities – suggest that appropriate model structure could vary spatially. Considering that improvements in weather forecasting have resulted in part from the increasingly high resolution of data and models [35,36], it is logical to expect that improvements in infectious disease forecasting will also require improvements in the spatial and temporal resolution of epidemiological data and models.

Our analysis identified intriguing differences in temporal incidence patterns across spatial scales, but at the same time there are important limitations to acknowledge. First, our conclusions are not dependent on the magnitude of transmission, but do require that patterns in case report data reflect patterns in underlying transmission. With a high rate of asymptomatic infection and the likelihood of extensive variability in reporting rates [37], particularly at the municipal level, much caution is due. Second, our ability to ascribe meaning to the groups identified by our classification algorithm was limited by the simplicity of our simulation model, particularly with respect to spatial interaction. Consequently, while this analysis identified important relationships between scale and epidemic characteristics, it does not provide a final or comprehensive understanding of the spatial transmission dynamics of ZIKV in Colombia. Third, our model relied on a simplified description of seasonal transmission, when in fact patterns of seasonality are likely to vary spatially and to interact strongly with introduction timing [38].

CONCLUSIONS

Previous analyses of Zika [8,27], as well as chikungunya [9,39], have drawn epidemiological inferences and made forecasts on the basis of nationally aggregated time series data. These efforts depend on the implicit assumption that spatially disaggregated temporal patterns are homogeneous and consistent with spatially aggregated temporal patterns. Our analysis showed that while national-level patterns may be somewhat reflective of departmental-level patterns, municipal-level patterns of cumulative incidence are exceedingly diverse and not well approximated by national-level patterns. Although our analysis was limited in its ability to explain the mechanisms that drove these diverse patterns, applying our classification algorithm to simulated data in which driving mechanisms were known showed that spatial differences in driving mechanisms can be associated with perceptible differences in temporal patterns. The initial wave of the Zika epidemic appears to have subsided, but understanding of spatial variation in transmission dynamics remains imperative for time-sensitive applications such as site selection for vaccine trials [40,41] and anticipating future epidemics [8].

LIST OF ABBREVIATIONS

CHIKV = chikungunya virus, ZIKV = Zika virus

DECLARATIONS

Ethics approval and consent to participate: Not applicable

Consent for publication: Not applicable

Availability of data and materials: The datasets and code used in the analysis will be made freely available online in a GitHub repository at the time of publication.

Competing interests: The authors declare that they have no competing interests.

Funding: This research was supported by a RAPID grant from the National Science Foundation (DEB 1641130) and by a Young Faculty Award from the Defense Advanced Research Projects Agency (D16AP00114).

Authors' contributions: TAP, IR-B, CM, CMB, MAJ, and RCR conceived of the research. IR-B, ASS, and GE prepared data sets. TAP, IR-B, CM, and RCR performed the analyses. TAP led the writing, and IR-B, CM, ASS, GE, CMB, MAJ, and RCR contributed to the writing and provided critical feedback to revisions.

Acknowledgements: Not applicable

REFERENCES

1. Turchin P, Taylor AD. Complex Dynamics in Ecological Time Series. Ecology. Ecological Society of America; 1992;73:289–305.

2. Hastings A. Timescales, dynamics, and ecological understanding. Ecology. Wiley Online Library; 2010;91:3471–80.

3. Bjørnstad ON, Grenfell BT. Noisy clockwork: time series analysis of population fluctuations in animals. Science. 2001;293:638–43.

4. Koelle K, Pascual M. Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. Am Nat. 2004;163:901–13.

5. Levin SA. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology. 1992;73:1943–67.

6. Salje H, Lessler J, Paul KK, Azman AS, Rahman MW, Rahman M, et al. How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. Proc Natl Acad Sci U S A. 2016;113:13420–5.

7. Salje H, Lessler J, Maljkovic Berry I, Melendrez MC, Endy T, Kalayanarooj S, et al. Dengue

diversity across spatial and temporal scales: Local structure and the effect of host population size. Science. 2017;355:1302–6.

8. Ferguson NM, Cucunubá ZM, Dorigatti I, Nedjati-Gilani GL, Donnelly CA, Basáñez M-G, et al. EPIDEMIOLOGY. Countering the Zika epidemic in Latin America. Science. 2016;353:353–4.

9. Escobar LE, Qiao H, Peterson AT. Forecasting Chikungunya spread in the Americas via data-driven empirical approaches. Parasit Vectors. 2016;9:112.

10. Instituto Nacional de Salud [Internet]. Boletín Epidemiológico. [cited 2017 May 6]. Available from: http://www.ins.gov.co/boletin-epidemiologico/Paginas/default.aspx

11. Chretien J-P, Rivers CM, Johansson MA. Make Data Sharing Routine to Prepare for Public Health Emergencies. PLoS Med. 2016;13:e1002109.

12. Boletín Epidemiológico - Todos los documentos [Internet]. [cited 2018 Feb 18]. Available from: http://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/Forms/AllItems.aspx

13. Sorichetta A, Hornby GM, Stevens FR, Gaughan AE, Linard C, Tatem AJ. High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. Sci Data. 2015;2:150045.

14. SIGOT : Sistema de información geográfica para la planeación y el ordenamiento territorial SIG-OT [Internet]. [cited 2018 Feb 18]. Available from: http://sigotn.igac.gov.co/sigotn/frames_pagina.aspx

15. Perkins TA, Siraj AS, Ruktanonchai CW. Model-based projections of Zika virus infections in childbearing women in the Americas. Microbiology [Internet]. nature.com; 2016; Available from: http://www.nature.com/articles/nmicrobiol2016126?WT.mc_id=EMX_NMB_1612_DecemberContent_Portfolio&WT.ec_id=EXTERNAL

16. Rojas DP, Dean NE, Yang Y, Kenah E, Quintero J, Tomasi S, et al. The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. Euro Surveill [Internet]. 2016;21. Available from: http://dx.doi.org/10.2807/1560-7917.ES.2016.21.28.30283

17. CRAN - Package EpiEstim [Internet]. [cited 2017 Dec 31]. Available from: https://cran.r-project.org/web/packages/EpiEstim

18. Cori A, Ferguson NM, Fraser C, Cauchemez S. A new framework and software to estimate time-varying reproduction numbers during epidemics. Am J Epidemiol. 2013;178:1505–12.

19. King AA, Domenech de Cellès M, Magpantay FMG, Rohani P. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. Proc Biol Sci. 2015;282:20150347.

20. Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. J Math Model Algorithms. Springer Netherlands; 2006;5:475–504.

21. “Finding Groups in Data”: Cluster Analysis Extended Rousseeuw et al. [R package cluster version 2.0.6]. Comprehensive R Archive Network (CRAN); [cited 2018 Jan 1]; Available from: https://cran.r-project.org/web/packages/cluster/index.html

22. Rousseeuw PJ. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.

23. Kucharski AJ, Funk S, Eggo RM, Mallet H-P, Edmunds WJ, Nilles EJ. Transmission Dynamics of Zika Virus in Island Populations: A Modelling Analysis of the 2013-14 French Polynesia Outbreak. PLoS Negl Trop Dis. 2016;10:e0004726.

24. Rand DA, Wilson HB. Using spatio-temporal chaos and intermediate-scale determinism to quantify spatially extended ecosystems. Proceedings of the Royal Society of London B: Biological Sciences. The Royal Society; 1995;259:111–7.

25. Keeling MJ, Mezić I, Hendry RJ, McGlade J, Rand DA. Characteristic length scales of spatial models in ecology via fluctuation analysis. Philos Trans R Soc Lond B Biol Sci. The Royal Society; 1997;352:1589–601.

26. Pascual M, Levin SA. FROM INDIVIDUALS TO POPULATION DENSITIES: SEARCHING FOR THE INTERMEDIATE SCALE OF NONTRIVIAL DETERMINISM. Ecology. Ecological Society of America; 1999;80:2225–36.

27. Shutt DP, Manore CA, Pankavich S, Porter AT, Del Valle SY. Estimating the reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America. Epidemics. 2017;21:63–79.

28. Mills HL, Riley S. The spatial resolution of epidemic peaks. PLoS Comput Biol. 2014;10:e1003561.

29. Durrett R, Levin S. The Importance of Being Discrete (and Spatial). Theor Popul Biol. 1994;46:363–94.

30. Grenfell BT, Bjørnstad ON, Kappey J. Travelling waves and spatial hierarchies in measles epidemics. Nature. 2001;414:716–23.

31. Cummings DAT, Irizarry RA, Huang NE, Endy TP, Nisalak A, Ungchusak K, et al. Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. Nature. 2004;427:344–7.

32. Dalziel BD, Pourbohloul B, Ellner SP. Human mobility patterns predict divergent epidemic dynamics among cities. Proc Biol Sci. 2013;280:20130763.

33. Rodriguez-Morales AJ, García-Loaiza CJ, Galindo-Marquez ML, Sabogal-Roman JA, Marin-Loaiza S, Lozada-Riascos CO, et al. Zika infection GIS-based mapping suggest high transmission activity in the border area of La Guajira, Colombia, a northeastern coast Caribbean department, 2015-2016: Implications for public health, migration and travel. Travel Med Infect Dis. 2016;14:286–8.

34. Scarpino SV, Petri G. On the predictability of infectious disease outbreaks [Internet]. arXiv [physics.soc-ph]. 2017. Available from: http://arxiv.org/abs/1703.07317

35. Mass CF, Ovens D, Westrick K, Colle BA. Does Increasing Horizontal Resolution Produce More Skillful Forecasts? Bull Am Meteorol Soc. American Meteorological Society; 2002;83:407–30.

36. Roebber PJ, Schultz DM, Colle BA, Stensrud DJ. Toward Improved Prediction: High-Resolution and Ensemble Modeling Systems in Operations. Weather Forecast. American

Meteorological Society; 2004;19:936–49.

37. Lessler J, Chaisson LH, Kucirka LM, Bi Q, Grantz K, Salje H, et al. Assessing the global threat from Zika virus. Science. 2016;353:aaf8160.

38. Huber JH, Childs ML, Caldwell JM, Mordecai EA. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission [Internet]. bioRxiv. 2017 [cited 2018 Feb 18]. p. 230383. Available from: https://www.biorxiv.org/content/early/2017/12/08/230383.abstract

39. Perkins TA, Metcalf CJE, Grenfell BT, Tatem AJ. Estimating drivers of autochthonous transmission of chikungunya virus in its invasion of the americas. PLoS Curr [Internet]. 2015;7. Available from: http://dx.doi.org/10.1371/currents.outbreaks.a4c7b6ac10e0420b1788c9767946d1fc

40. Perkins TA. Retracing Zika's footsteps across the Americas with computational modeling. Proc Natl Acad Sci U S A. National Academy of Sciences; 2017;114:5558–60.

41. Asher J, Barker C, Chen G, Cummings D, Chinazzi M, Daniel-Wayman S, et al. Preliminary results of models to predict areas in the Americas with increased likelihood of Zika virus transmission in 2017 [Internet]. bioRxiv. 2017 [cited 2018 Feb 18]. p. 187591. Available from: https://www.biorxiv.org/content/early/2017/09/29/187591

SUPPORTING INFORMATION

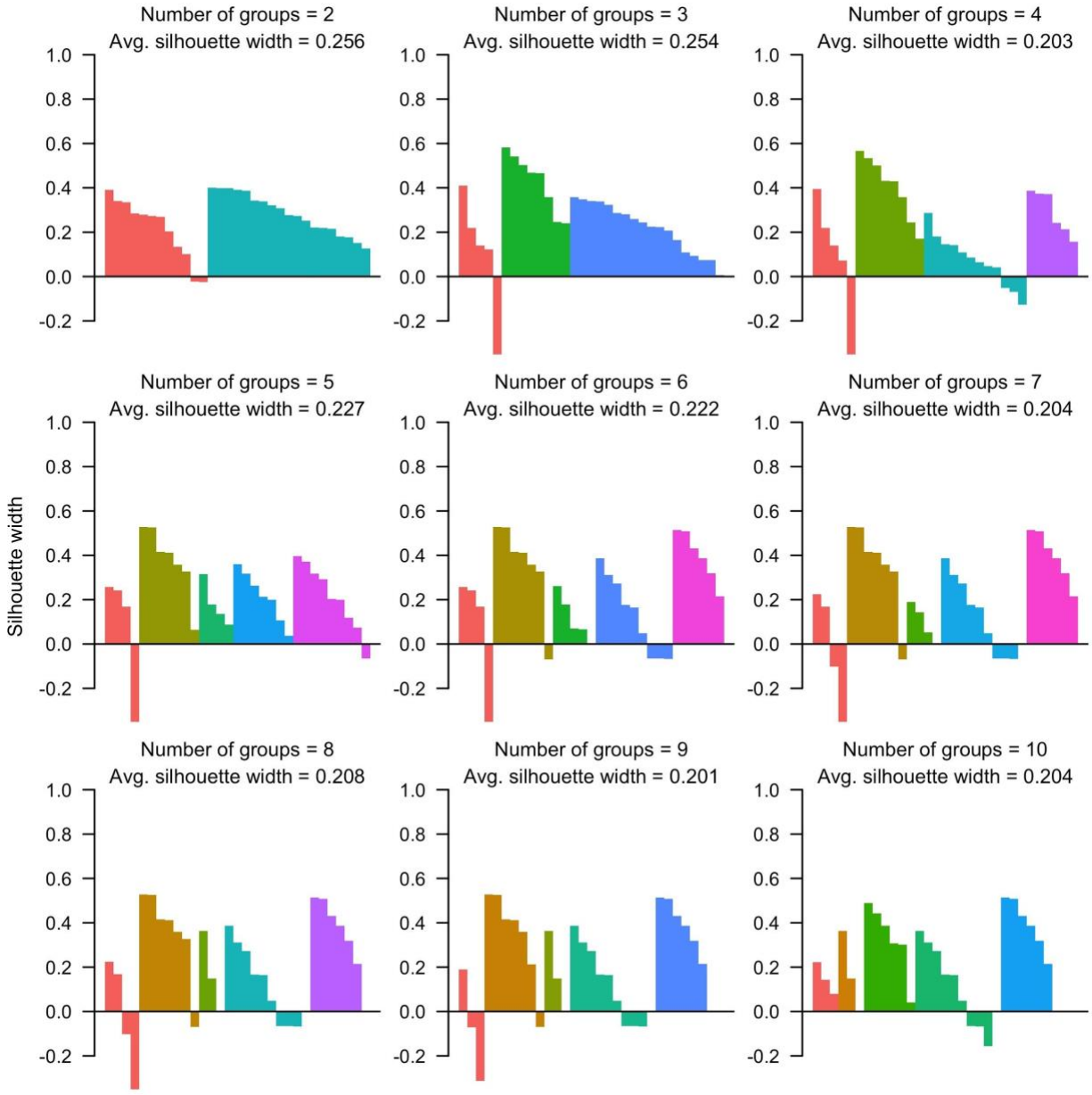

Figure S1. Silhouette plots at the departmental level for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given department according to the group assignments indicated by different colors in each panel. Higher average silhouette values indicate stronger clustering.

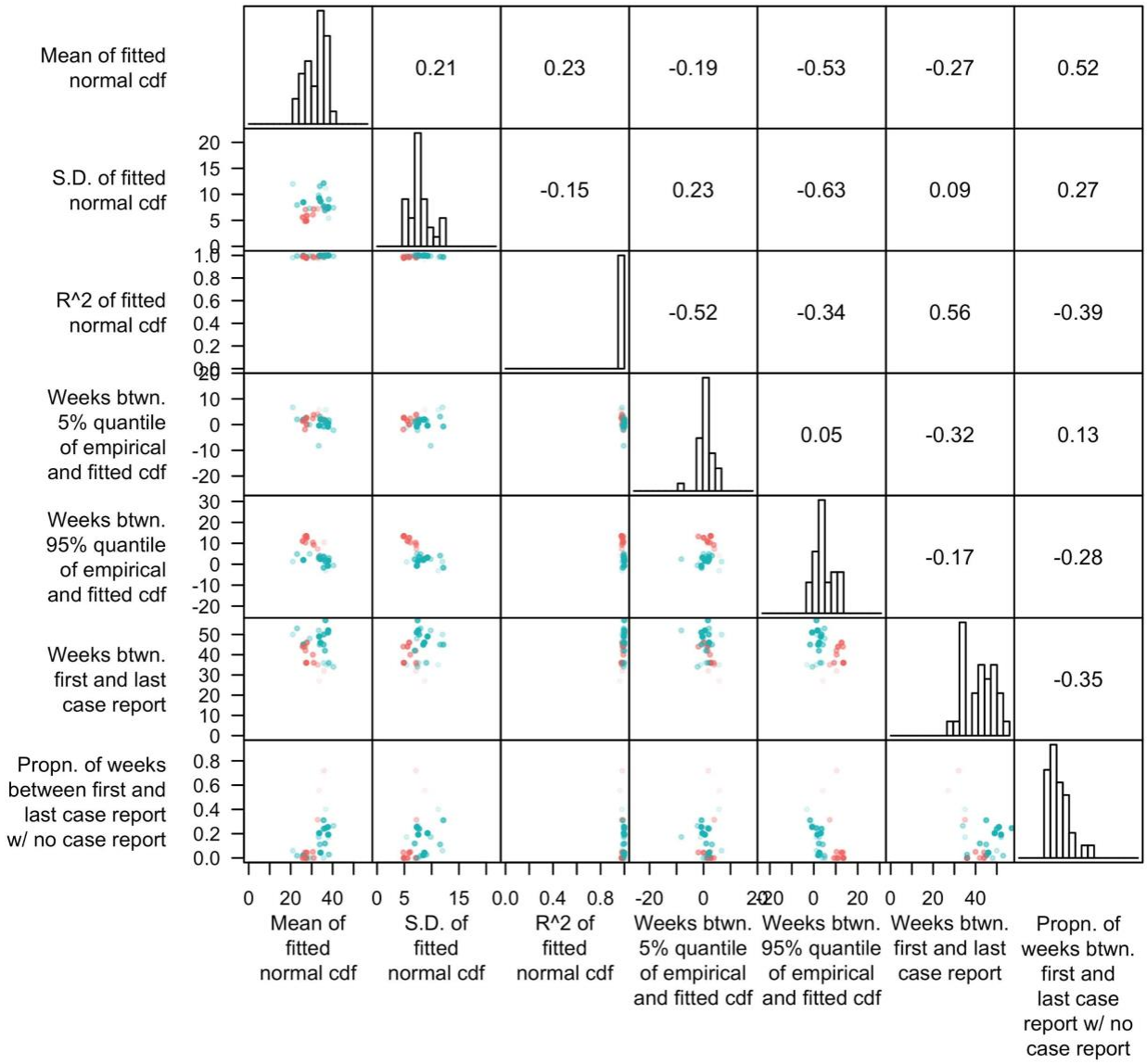

Figure S2. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the departments into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

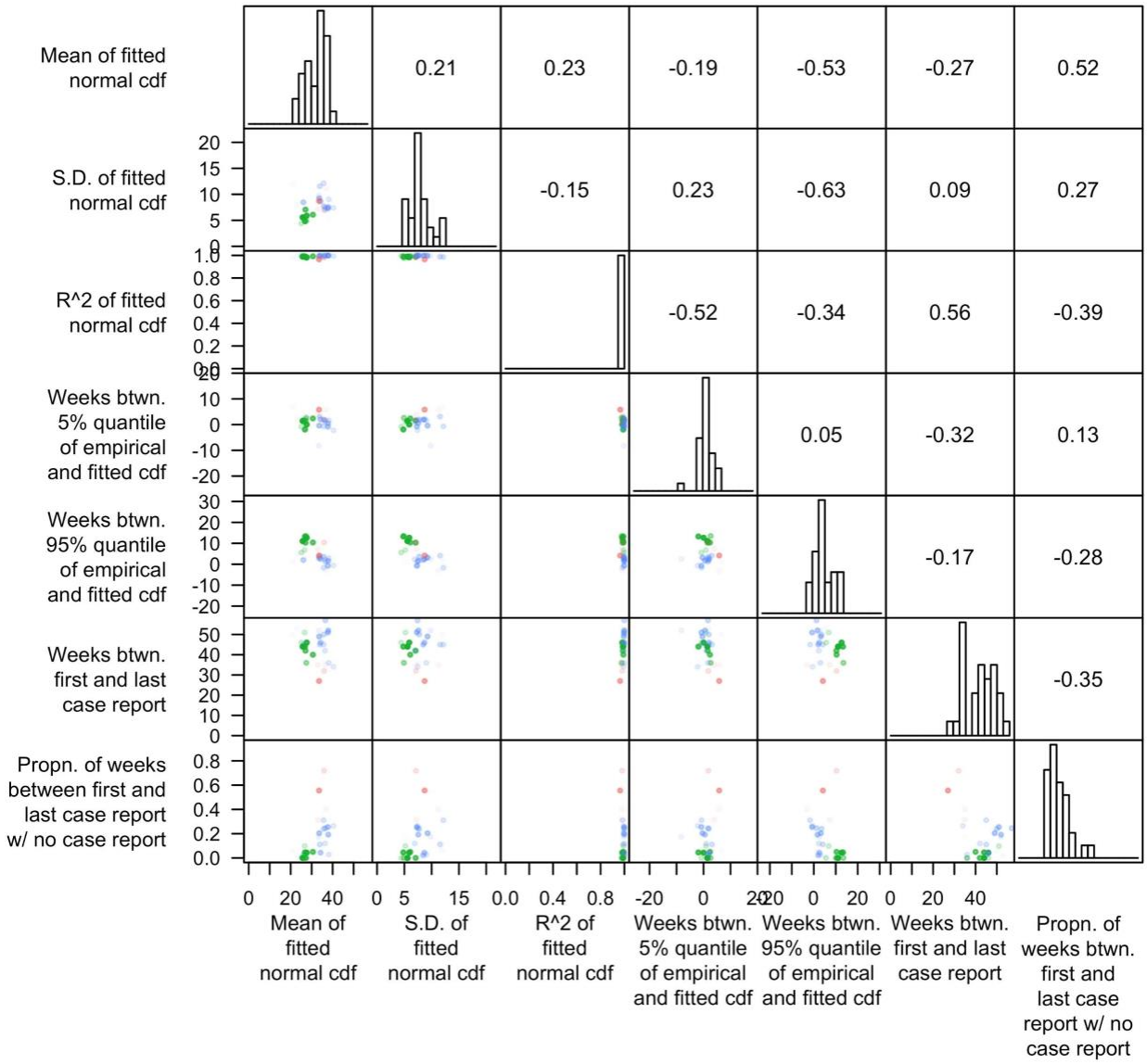

Figure S3. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the departments into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

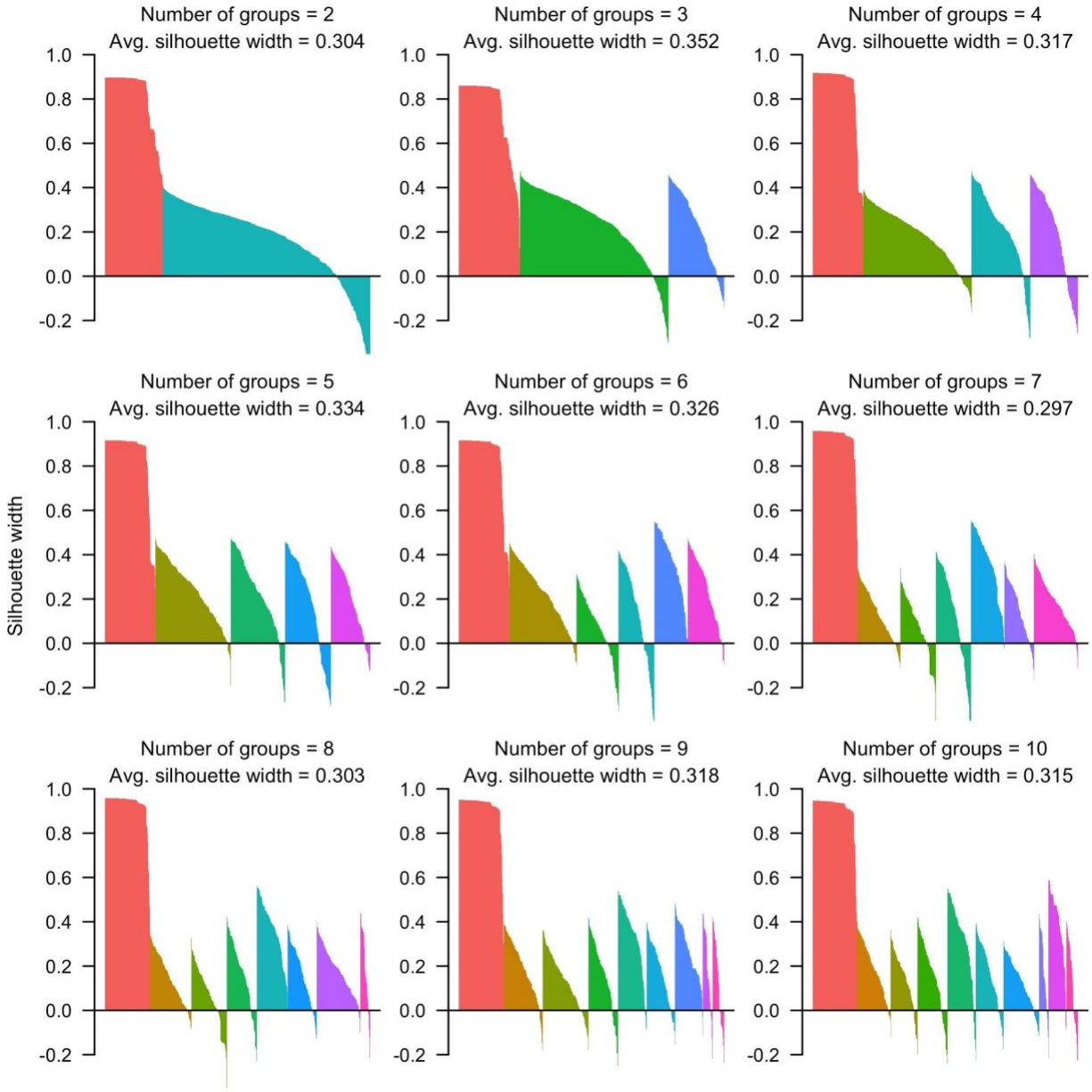

Figure S4. Silhouette plots at the municipal level for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given municipality according to the group assignments indicated by different colors in each panel. Higher average silhouette values indicate stronger clustering.

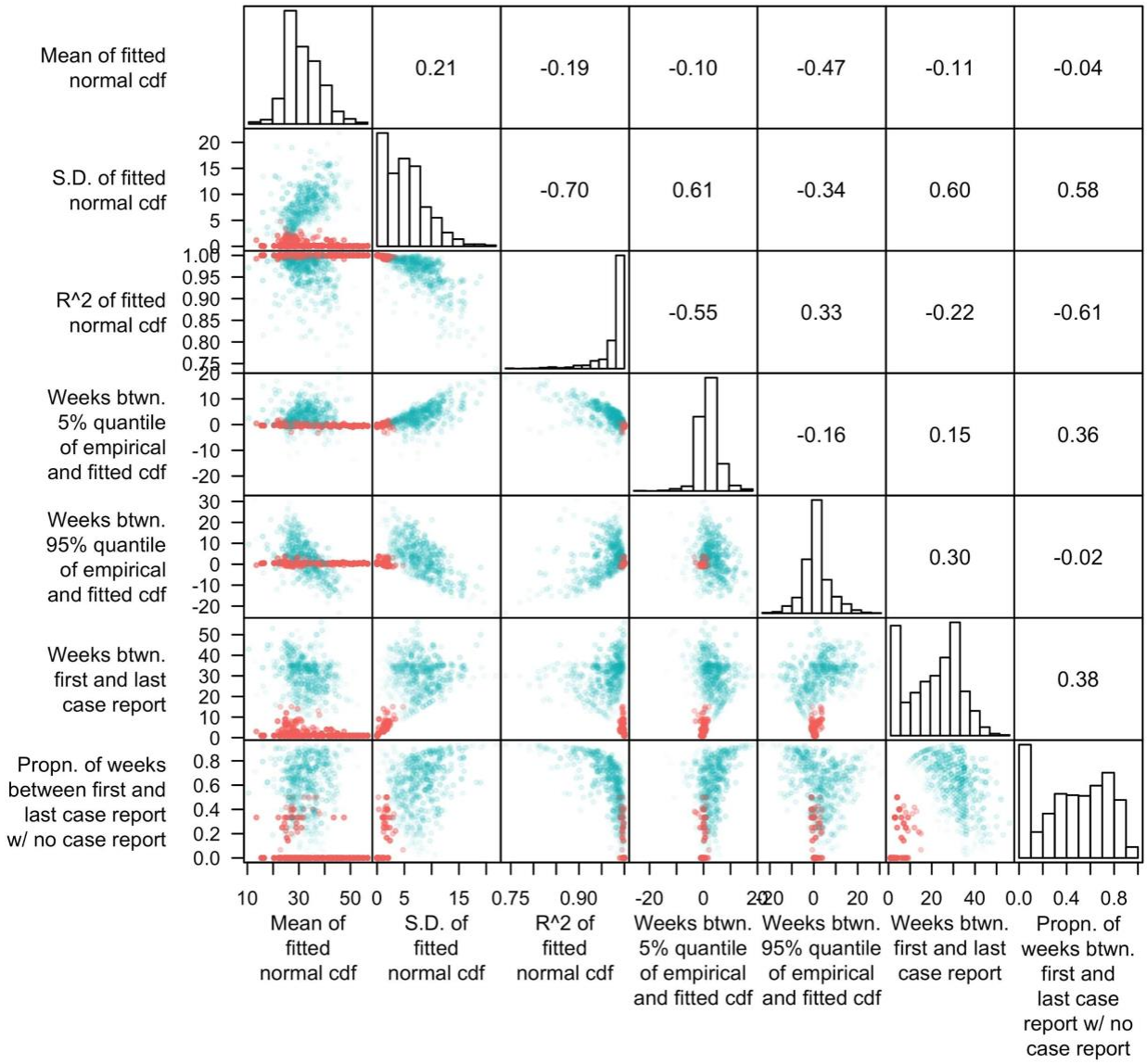

Figure S5. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

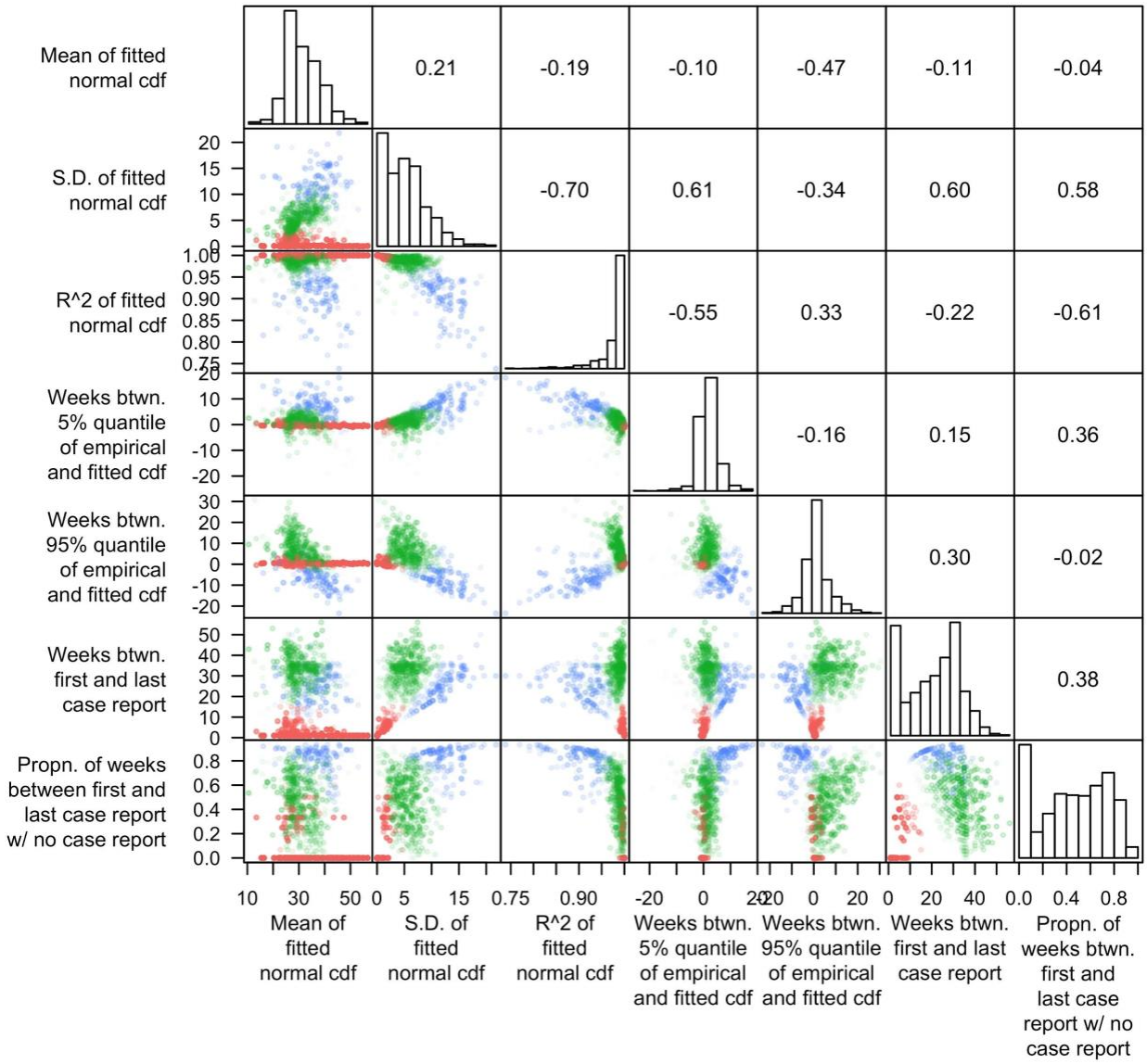

Figure S6. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

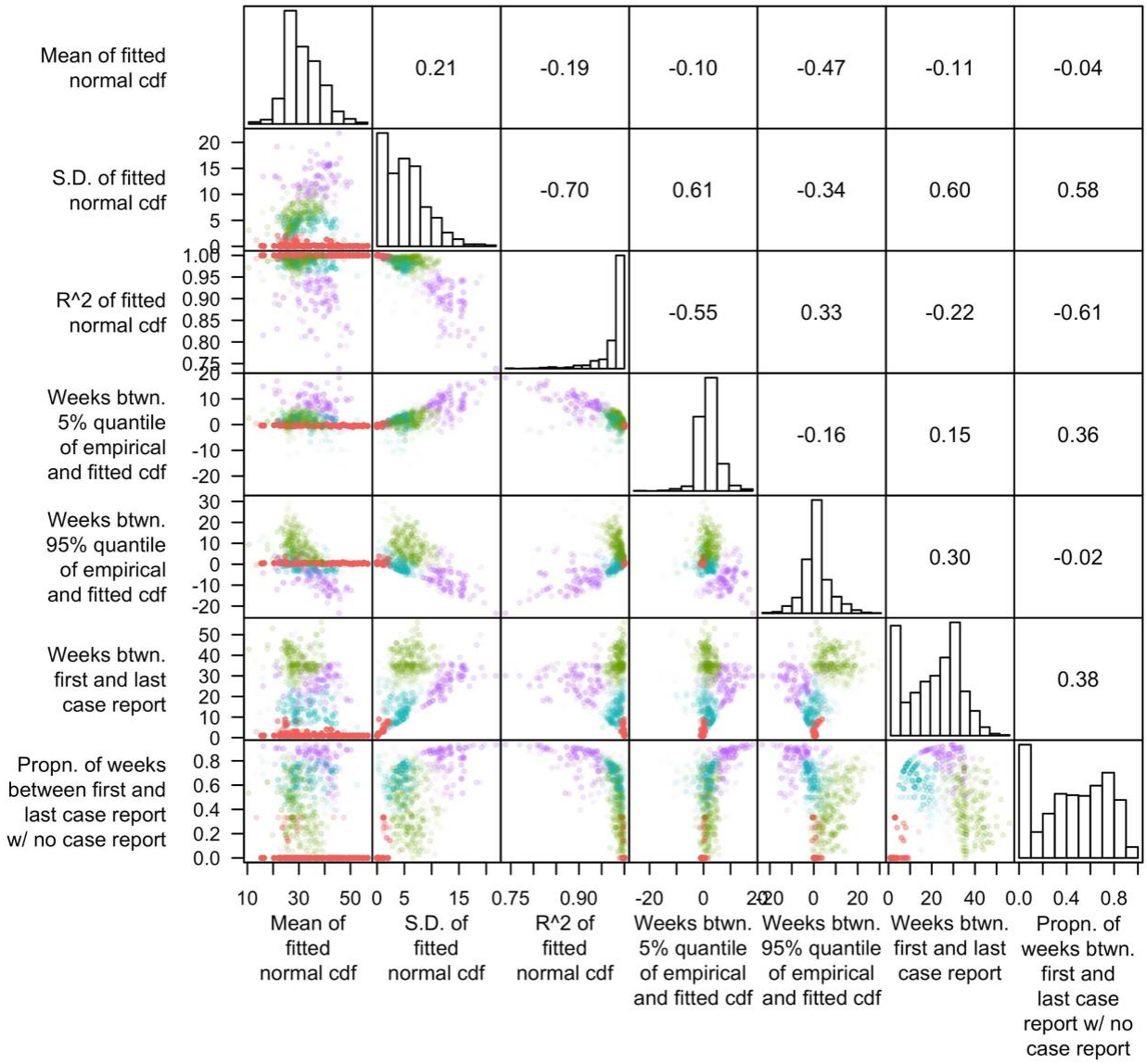

Figure S7. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of four groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

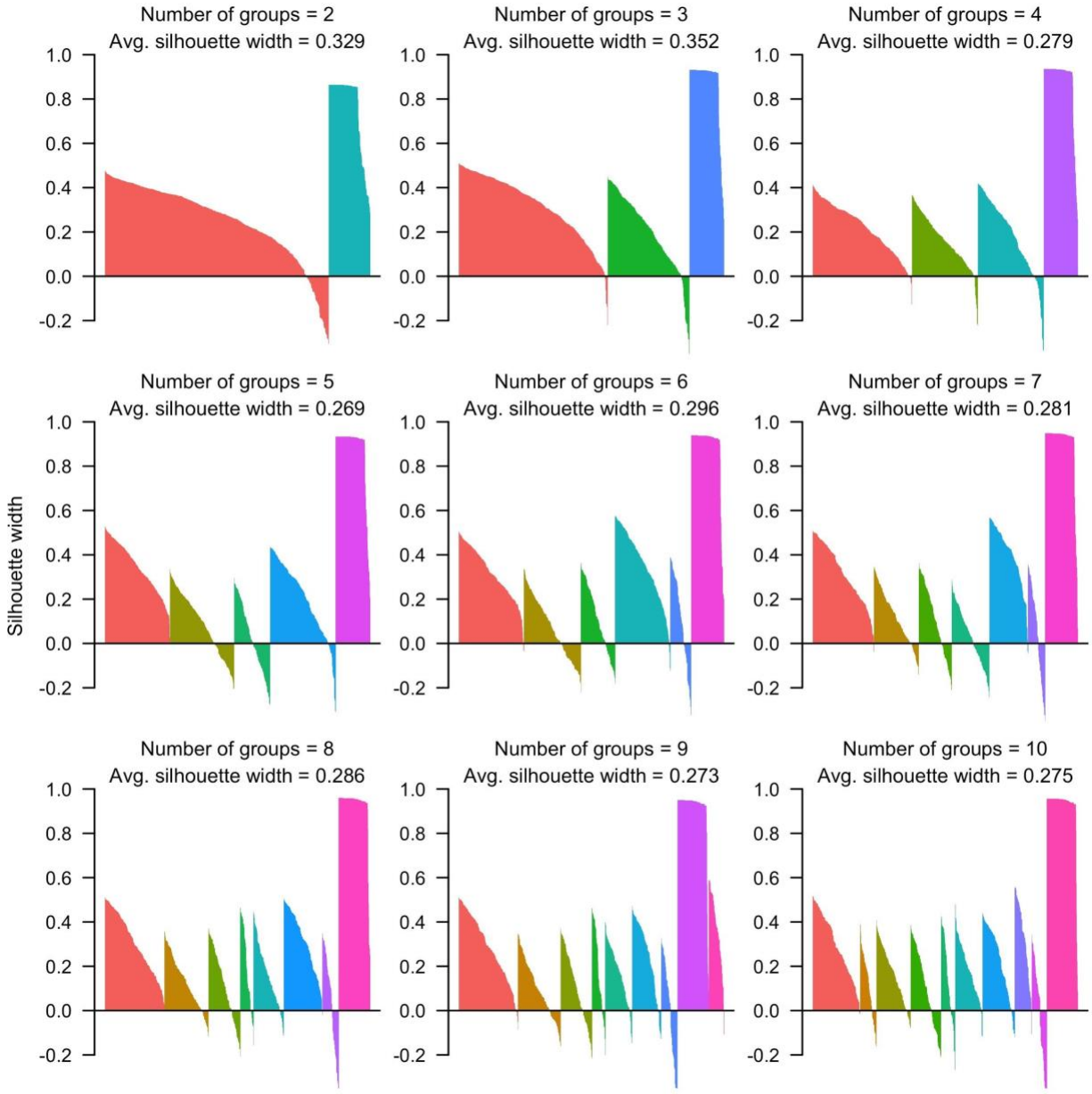

Figure S8. Silhouette plots at the municipal level based on a randomly selected simulated data set for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given municipality according to the group assignments indicated by different colors in each panel. Higher average silhouette values indicate stronger clustering.

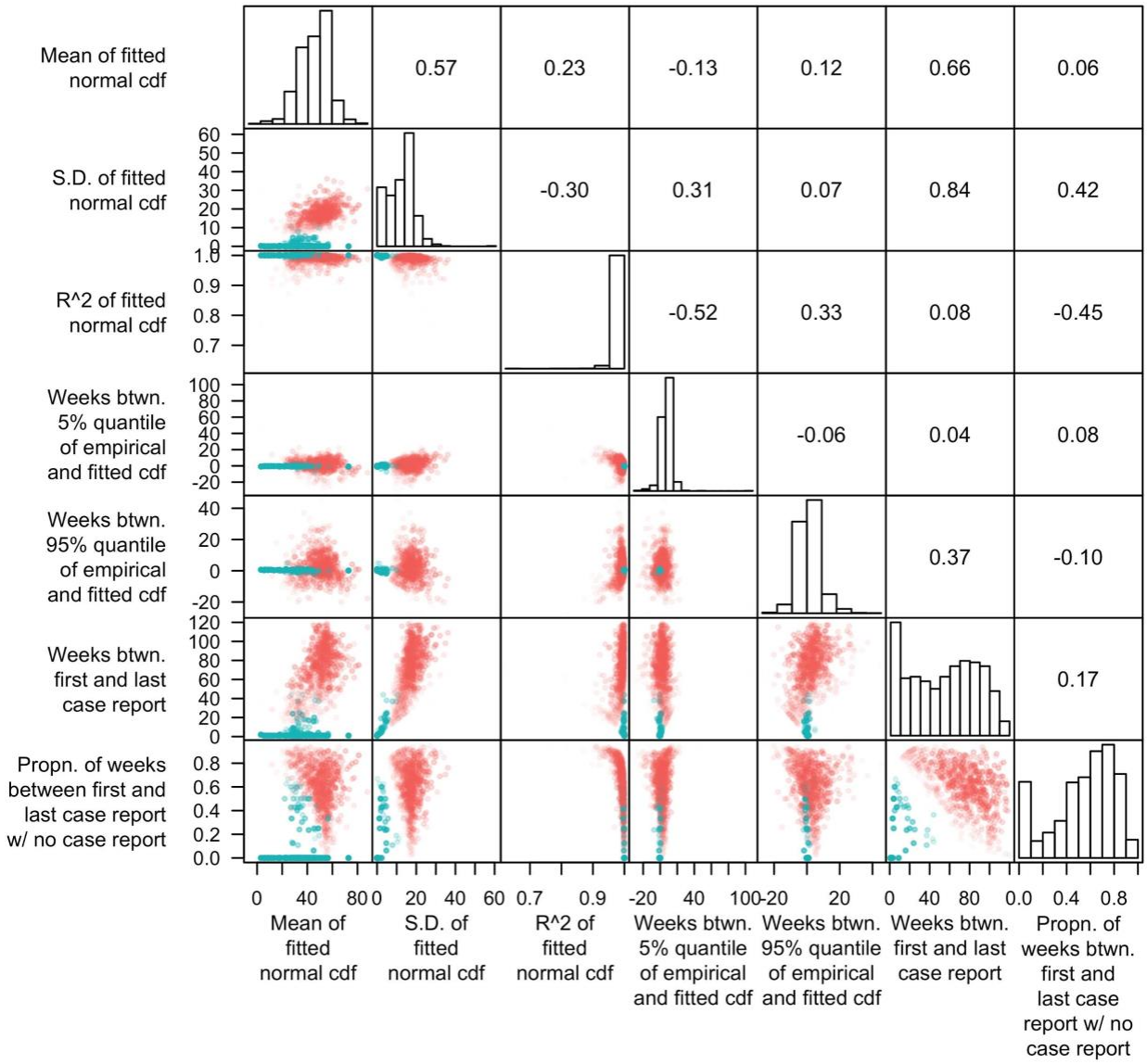

Figure S9. Pairwise plots of features of proportional cumulative incidence curves based on a randomly selected simulated data set, with colors distinguishing group assignment of the municipalities into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

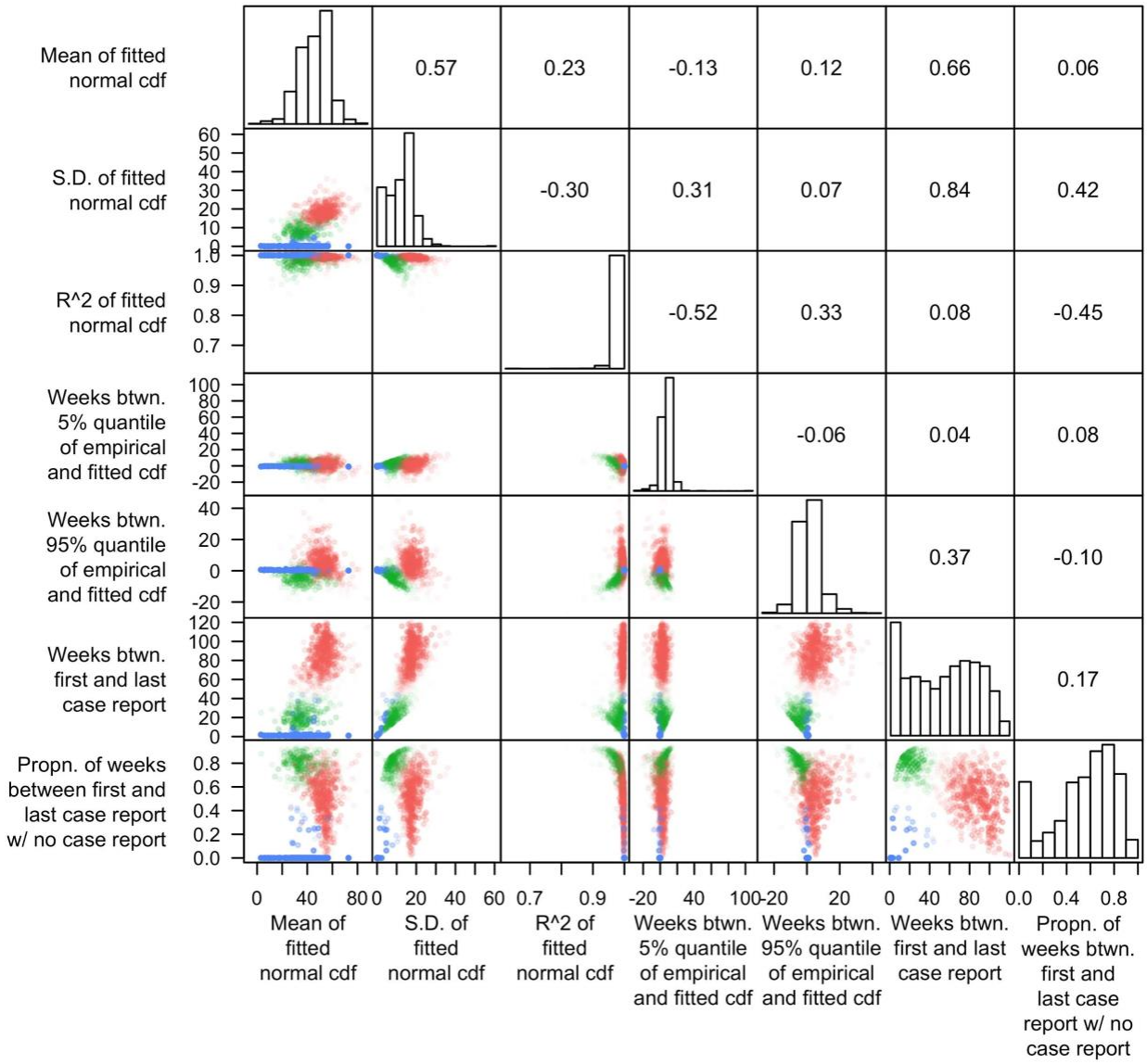

Figure S10. Pairwise plots of features of proportional cumulative incidence curves based on a randomly selected simulated data set, with colors distinguishing group assignment of the municipalities into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.