Heterogeneous local dynamics revealed by classification analysis of spatially disaggregated time series data

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, taperkins@nd.edu, asiraj@nd.edu, guido.espana@nd.edu
[2] Department of Medicine, University of California, San Francisco, Isabel.Rodriguez@ucsf.edu
[3] Theoretical Biology and Biophysics, Los Alamos National Laboratory, cbearmath@gmail.com
[4] Department of Pathology, Microbiology, and Immunology, University of California, Davis, cmbarker@ucdavis.edu
[5] Division of Vector-Borne Diseases, Centers for Disease Control and Prevention, mjohansson@cdc.gov
[6] Center for Communicable Disease Dynamics, Harvard T.H. Chan School of Public Health
[7] Institute for Health Metrics and Evaluation, University of Washington, bcreiner@uw.edu
* Contributed equally

Correspondence:
Alex Perkins, 100 Galvin Hall, Notre Dame, IN 46556; tel: 574-631-7179; fax: 574-631-7413; email: taperkins@nd.edu
Robert Reiner, 325 9th Avenue, Box 359931, Seattle, WA 98104; tel: 206-744-8493; fax: 206-744-3693; email: bcreiner@uw.edu

ABSTRACT

Time series data provide a crucial window into ecological dynamics, yet their utility is often limited by the spatially aggregated form in which they are presented. When working with time series data, violating the implicit assumption of homogeneous dynamics below the scale of aggregation could bias inferences about underlying processes. We tested this assumption in the context of the 2015-2016 Zika epidemic in Colombia, where time series of weekly case reports were available at national, departmental, and municipal scales. First, we performed a descriptive analysis, which showed that the timing of departmental-level epidemic peaks varied by three months and that departmental-level estimates of the time-varying reproduction number, $R(t)$, showed patterns that were distinct from a national-level estimate. Second, we applied a classification algorithm to six features of cumulative incidence curves, which showed that variability in epidemic duration, the length of the epidemic tail, and consistency with a normal distribution function made the greatest contributions to distinguishing groups. Third, we applied this classification algorithm to data simulated with a stochastic transmission model, which showed that group assignments were consistent with simulated differences in the basic reproduction number, R_0. This result, along with associations between spatial drivers of transmission and group assignments based on observed data, suggests that the classification algorithm is capable of detecting differences in temporal patterns that are associated with differences in underlying ecological drivers. Overall, this diversity of temporal patterns at local scales underscores the value of spatially disaggregated time series data.

INTRODUCTION

Time series have been used for many years to make inferences about processes that shape the dynamics of a wide range of ecological systems (Turchin & Taylor 1992). This long history has resulted in appreciation of a number of common challenges for time series analysis (Hastings 2010). 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 density-dependent feedbacks (Bjørnstad & Grenfell 2001; Koelle & Pascual 2004). An even more fundamental challenge lies in defining the time series in the first place, especially with respect to space (Levin 1992). The question is, at what spatial scale should ecological 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 (Salje *et al.* 2016, 2017), applications ranging from estimation of time-varying reproduction numbers (Ferguson *et al.* 2016) to forecasting (Escobar *et al.* 2016, Del Valle *et al.* 2018) have utilized data aggregated at national scales for countries as vast and spatially heterogeneous as Brazil and Mexico.

Unlike most other countries in the Americas, routine surveillance of Zika in Colombia was reported on a weekly basis in each of its 1,123 municipalities during the 2015-2016 epidemic (INS 2017). Although such 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 most countries (Chretien *et al.* 2016). 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 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 of ZIKV transmission to determine the extent to which distinct temporal patterns may reflect distinct ecological drivers. All data and code used in this study are available at https://github.com/TAlexPerkins/TimeSeriesSpatialScale.

METHODS

Data

The focal point of our analysis was a collection of municipal-level time series of weekly Zika case reports at the municipal level in Colombia spanning August 2015 through September 2016. The primary source of these data was the Colombian National Institute of Health (Instituto Nacional de Salud, INS), which made official weekly reports of the cumulative numbers of suspected and confirmed Zika cases available in real time during the epidemic (Boletín 2018). The version of these data that we used in this analysis were processed in a manner that addressed inconsistencies between data reported at municipal and departmental scales, as described by Siraj *et al.* (2018). Specifically, to correct for the fact that the total of municipal-level data from 2015 (3,875 cases) was less than the total of national-level data from 2015 (11,712), we imputed the 7,837 missing cases at the municipal level for 2015 by multiplying each municipality's weekly incidence in 2015 by a factor required to achieve better known cumulative totals for each municipality as of the first week of 2016.

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 bar plot of national case reports color-coded by which of 33 departments those national cases arose from. Likewise, for each of those departments, we generated a bar plot 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.* (2016), we used the EstimateR function from the EpiEstim library (Cori 2013) in R to estimate $R(t)$ for each time series based on the method introduced by Cori *et al.* (2013). In brief, this method is based on an assumed distribution of the serial interval (i.e., the timing between onset of primary and secondary cases) that can be used to estimate the number of cases in the previous generation that gave rise to those observed in the present generation, thereby enabling estimation of $R(t)$.

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 would 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 (King *et al.* 2015) 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 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

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

We excluded 2/33 departments and 307/1,123 municipalities 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:

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

Four of these features were defined 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.

We explored variation in $C_{i,t}$ at both departmental and municipal scales. To describe how variation in $C_{i,t}$ curves at those scales was distributed across the six-dimensional feature space, we performed a partitioning around medoids (PAM) clustering analysis (Reynolds *et al.* 2006) on centered and scaled values of the features using the pam function in the cluster library (Maechler et al. 2017) 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 one point is from points in the next most similar group compared to points in its own group (Rousseeuw 1987). 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 well with the group to which they are assigned.

Elucidation of driving processes

To aid in the interpretation of the classification analysis of empirical patterns of temporal incidence, we performed identical analyses of simulated patterns of temporal incidence. The value of doing so is that it provides a form of validation of the classification analysis: i.e., demonstrating that it is capable of identifying groups that correspond to known differences in underlying ecological processes. For this analysis, we defined groups of municipalities on the basis of whether the simulated R_0 value for a municipality was above or below 1, given the significance of this threshold for determining invasion outcomes. We performed classification analyses on 100 data sets simulated with a stochastic model of ZIKV transmission developed by Ferguson *et al.* (2016) and tailored to Colombia as described in Appendix S1. This particular model was chosen with the goal of simulating cumulative incidence curves at the municipality level that would be plausible with respect to the six features used in the classification analysis. As a result, the model has limitations that mean that it may not be appropriate for analyses of other, more complicated aspects of spatiotemporal transmission dynamics.

Although the analysis of simulated data provides a test of the algorithm, it does not facilitate inference of whether there truly are differences in ecological processes underlying the empirical data. Doing so convincingly would require more comprehensive analyses, ideally involving data about variables assumed to play an intermediary role in a hypothesized causal pathway between environmental variables and disease incidence (Metcalf *et al*. 2017). To explore whether there might at least be perceptible associations between ecological factors and groups identified by the classification analysis, we performed a series of one-way analyses of variance at both departmental and municipal scales. Specifically, our objective was to examine whether mean values of relevant environmental variables differed across these groups. Variables that we examined included R_0 values derived from Perkins *et al*. (2016) as described in Appendix S1, and seven variables compiled for municipalities and departments in Colombia by Siraj *et al*. (2018): *Ae. aegypti* occurrence probability, two measures of normalized difference vegetation index (NDVI), mean temperature, percent urban land cover, human population, and the gross cell product (GCP), a spatially disaggregated version of the gross domestic product economic index.

RESULTS

Descriptive analysis of weekly case reports

As a whole, the temporal pattern at the national level was consistent with what could be construed as 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 pattern can be used to estimate the temporal trajectory of the effective reproduction number, $R(t)$ (Cori *et al.* 2013). 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.

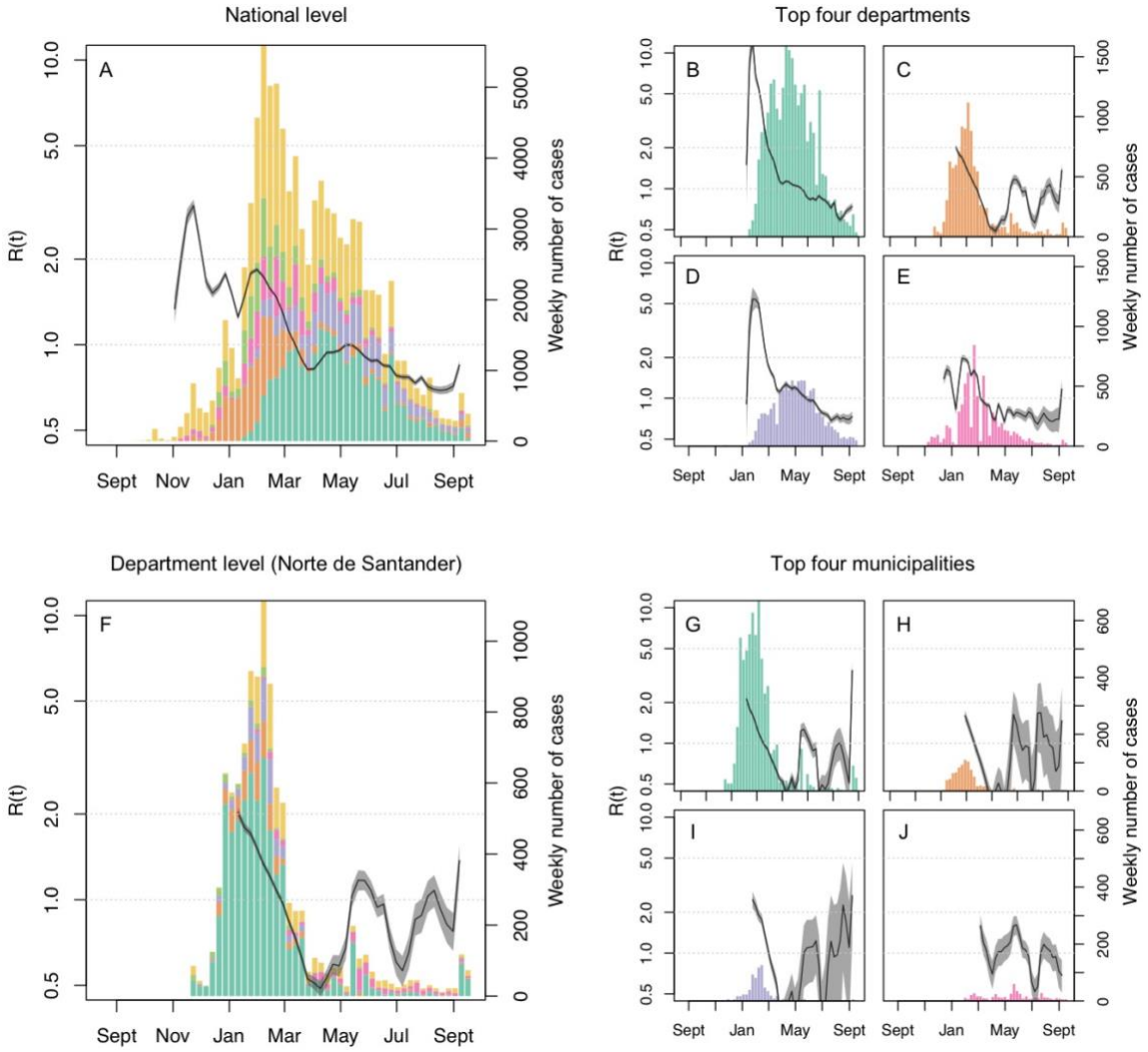

Figure 1. Weekly Zika case reports at the national level (A), for each of the four departments with the largest case report totals (B: Valle del Cauca; C: Norte de Santander; D: Santander; E: Tolima), at the departmental level for Norte de Santander (F), and for each of its four municipalities with the largest case report totals (G: Cucuta; H: Villa del Rosario; I: Los Patios; J: Ocaña). On the top row, colors match across A and B-E, with the addition of yellow in A that includes all departments other than those in B-E. On the bottom row, colors match across F and G-J, with the addition of yellow in F that includes all municipalities other than those in G-J. 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 those 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 1 (Fig. 1C).

Examination of temporal patterns at the municipal scale revealed even more variability in temporal 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. 1F). 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. 1G & 1H). Other municipalities in the department had only low, erratic incidence with no sign of a distinct epidemic (e.g., Fig. 1J). 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.

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. 2, top 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. 2, top 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. 2, top) but were generally apparent for the curves associated with the groups as a whole (Fig. S4). Spatially, groups tended to cluster along northern, central, and southern strata (Fig. 3, left), with incidence-weighted cartographs showing that the epidemic was mostly dominated by distinct northern and central strata (Fig. 3, top right).

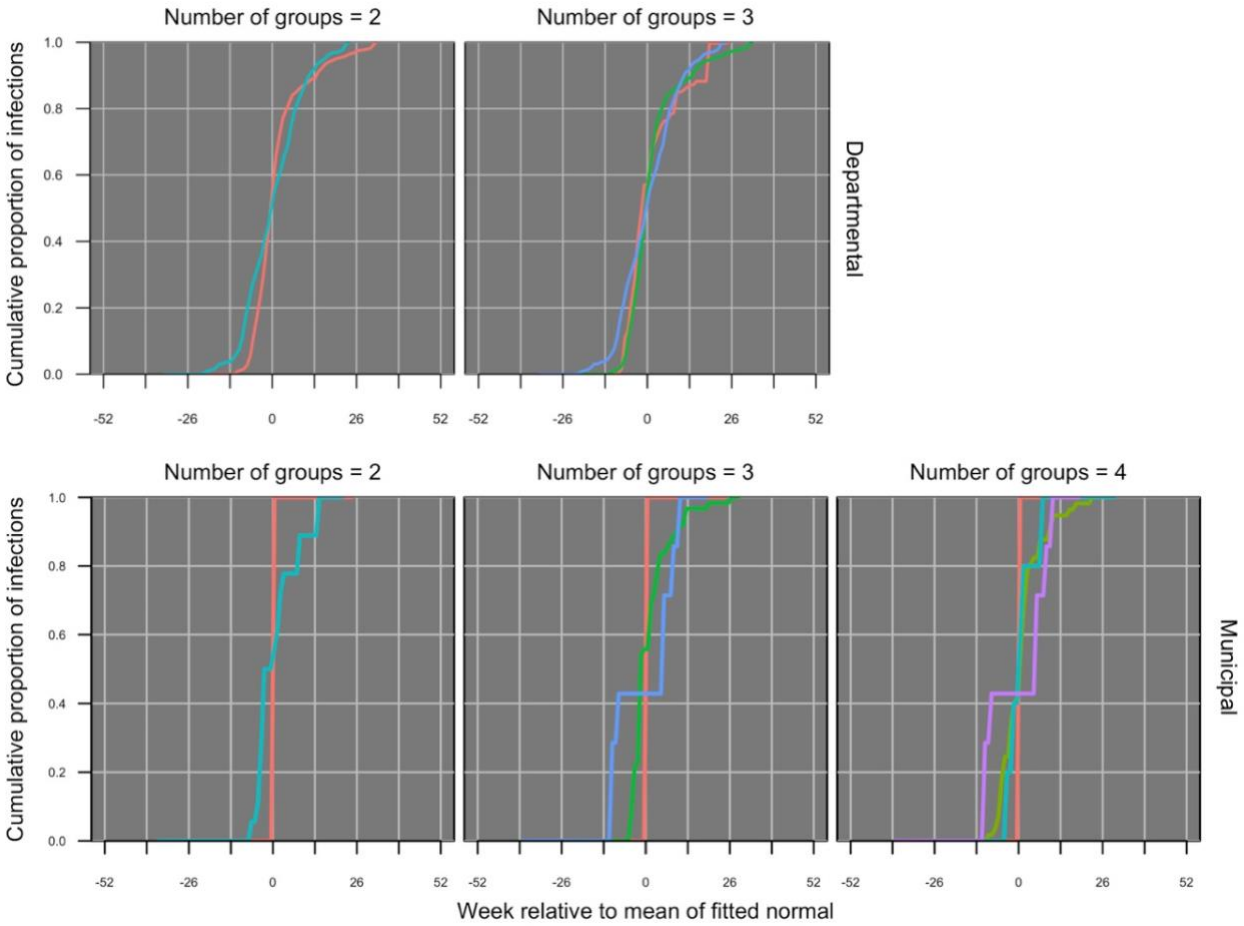

Figure 2. Proportional cumulative incidence curves at the departmental level (top) with two (left) or three (right) groups and at the municipal level (bottom) with two (left), three (middle), and four (right) groups. Only one representative curve is shown for each group, with that curve being chosen on the basis of being associated with the medoid of its group.

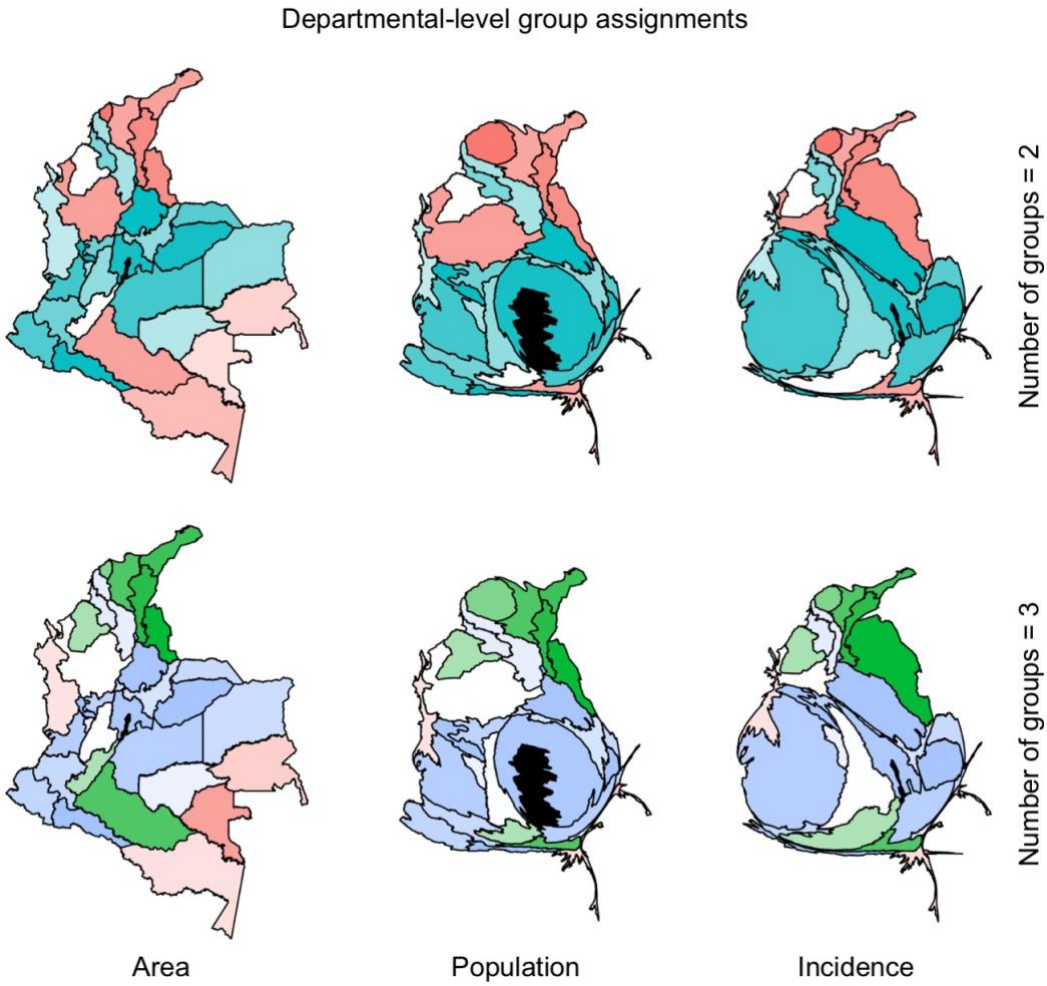

Figure 3. Cartograms at the departmental level weighted by area (left), population (center), and incidence (right). Department assignments to two (top) and three (bottom) groups are indicated by color, with transparency inversely proportional to silhouette value. The one department (Bogotá) with zero 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. S5). $F_{\Delta t}$ and F_{SD} were the features that were most important in distinguishing two groups (Fig. S6), $F_{95\%}$ made additional contributions to distinguishing three groups (Fig. S7), and F_{R^2} contributed to distinguishing four groups (Fig. S8). 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. 2, bottom). Some differences among the other groups were also apparent in the proportional cumulative incidence curves, with some having a long tail (Fig. 2, bottom middle: green) or two discrete jumps (Fig. 2, bottom 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 (compare Fig. 2 bottom with Fig. S9). 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. 4). 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. 4, 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. 4, right column).

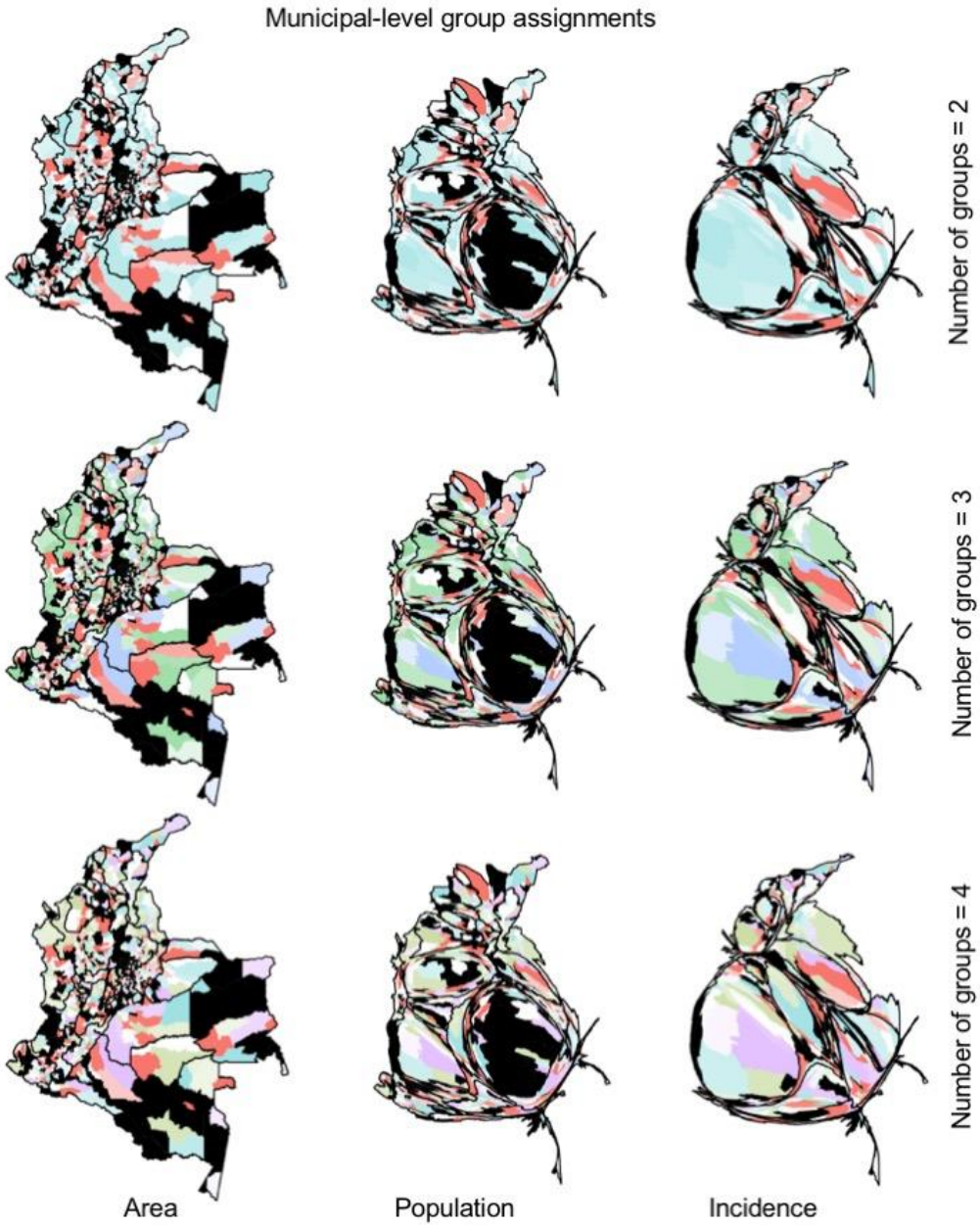

Figure 4. Cartograms at the municipal level weighted by area (left), population (center), and incidence (right). Municipality assignments to two (top), three (middle), and four (bottom) groups are indicated by color, with transparency inversely proportional to silhouette value. Municipalities with zero 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.

Elucidation of 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. S10 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. S11 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. S12 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. 5, 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. 5, 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. 6).

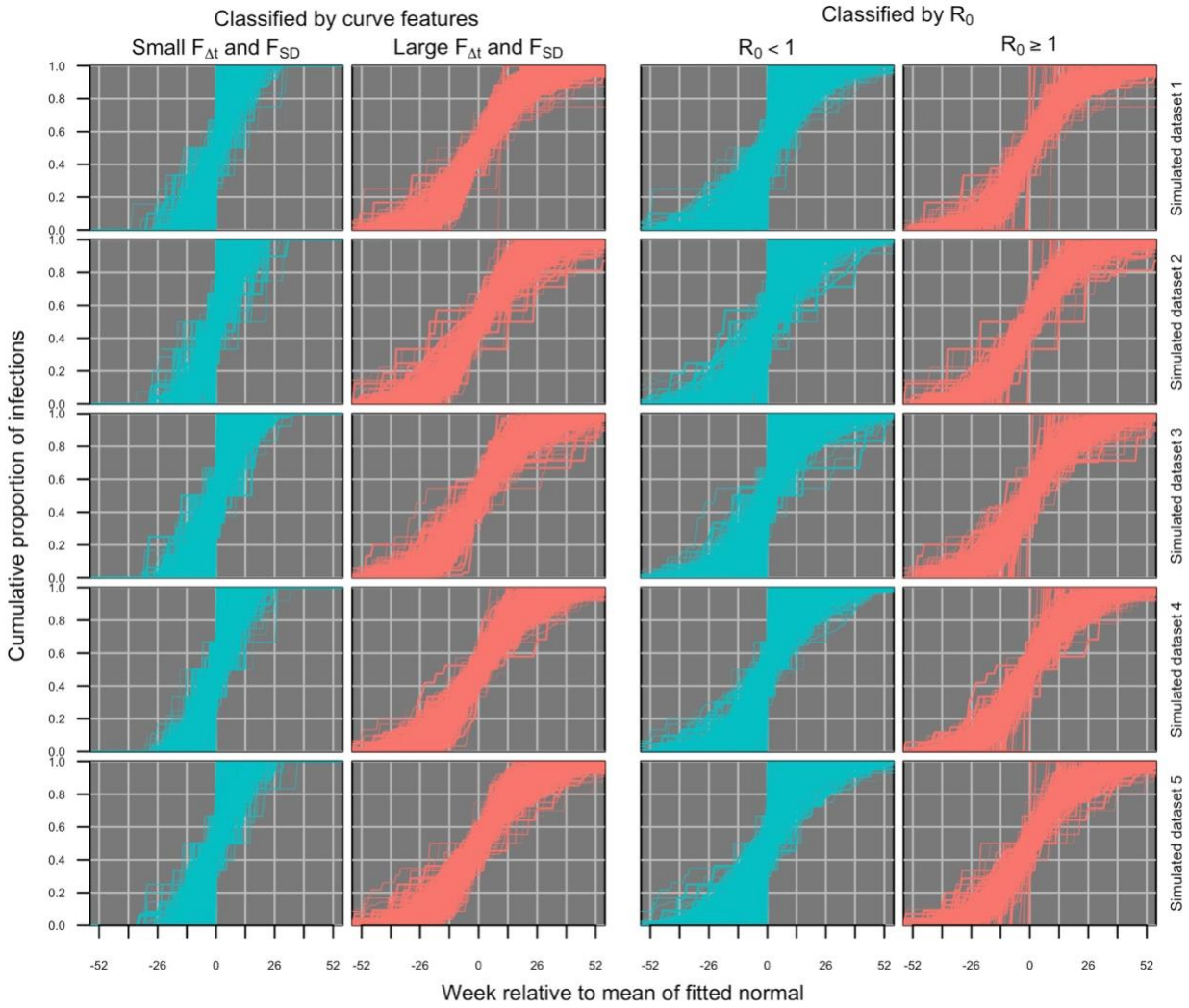

Figure 5. Proportional cumulative incidence curves at the municipal level from five randomly selected simulated datasets. The left two columns show two different groups classified by the curve classification algorithm, and the right two columns show two different groups defined by whether those municipalities have a R_0 above or below 1.

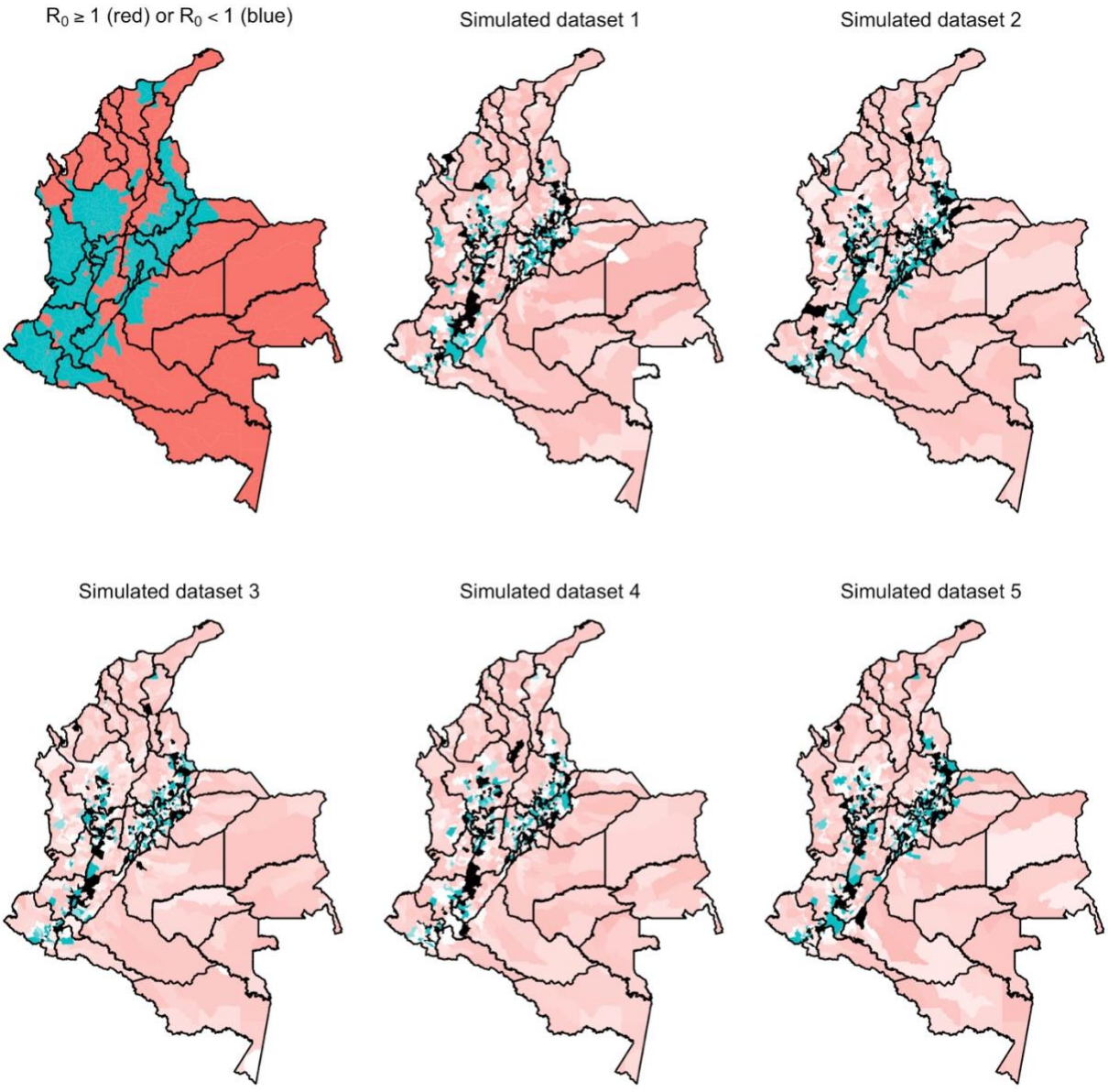

Figure 6. Cartograms at the municipal level weighted by area based on five randomly selected simulated datasets. Each municipality's status as having $R_0 > 1$ (red) or $R_0 < 1$ (blue) is indicated in the top left panel. In each of five simulated datasets shown in the other panels, municipality assignments to two groups are indicated by color, with transparency inversely proportional to silhouette value.

With respect to the empirical data, group assignments at the municipal scale were associated with perceptible differences in relevant environmental variables. For two groups, differences between groups were statistically significant for all eight variables examined ($p<0.002$ for all; Table S1). The group typified by steep, short curves (Fig. 2, bottom left: red), was associated with lower *Ae. aegypti* occurrence probability (0.04 vs. 0.05; $F_{836}=19.7$, $p<10^{-5}$), higher NDVI (aqua: 0.09 vs. 0.07; $F_{836}=12.9$, $p<10^{-3}$) (terra: 0.10 vs. 0.07; $F_{836}=13.3$, $p<10^{-3}$), lower temperature (21.7 vs. 23.9 °C; $F_{836}=32.9$, $p<10^{-7}$), lower urban cover (0.02 vs. 0.07;

F_{836}=29.6, $p<10^{-7}$), lower population (13,506 vs 49,144; F_{836}=10.2, $p<10^{-2}$), lower GCP (6,016 vs. 6,676; F_{836}=8.3, $p<10^{-2}$), and lower R_0 (1.1 vs. 1.7; F_{836}=16.1, $p<10^{-4}$) (Table S1). Differences among groups were significant only for the urban cover variable for three groups, and for no variables for four groups (Table S1). At the departmental scale, group assignments based on empirical data were generally not associated with differences in relevant environmental variables (Table S2).

DISCUSSION

Temporal incidence patterns play a vital role in inferring ecological dynamics and drivers thereof. By analyzing data from the 2015-2016 Zika epidemic in Colombia, we showed that temporal patterns can appear very different depending on the spatial scale at which data are aggregated. Whereas national-level dynamics appeared to follow a unimodal pattern consistent with behavior of standard epidemic models, departmental-level dynamics were somewhat more varied and municipal-level dynamics 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 patterns subnationally and that much of that variation may be driven by spatial variation in local conditions. Associations between group assignments and relevant environmental variables were most apparent at the municipal scale, consistent with the hypothesis that linkages between temporal dynamics and underlying ecological processes are strongest at fine spatial scales.

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 (Rand & Wilson 1995). 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 (Keeling *et al.* 1997; Pascual & Levin 1999). Epidemics, however, are inherently transient in nature, 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 (Shutt *et al.* 2017). 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 (Mills & Riley 2014).

Both stochasticity and spatial interaction are expected to contribute to variability in temporal dynamics at local scales (Durrett & Levin 1994). 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 about spatial interaction 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 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 (Grenfell *et al.* 2001; Cummings *et al.* 2004; Dalziel *et al.* 2013; Rodriguez-Morales *et al.* 2016). Future work that includes more realistic spatial interaction among subnational units would be helpful for resolving the hypothesis proposed here about the importance of spatial interaction in shaping temporal patterns at each of the spatial scales that we considered.

Our analysis identified intriguing differences in temporal patterns across spatial scales, but at the same time there are important limitations to acknowledge. First, although our conclusions are not dependent on the magnitude of transmission, they 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 (Lessler *et al.* 2016), particularly at the municipal level, some 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 spatial scale and epidemic characteristics, it does not provide a complete 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 could vary spatially and interact strongly with introduction timing (Huber *et al.* 2017).

Previous analyses of Zika (Ferguson *et al.* 2016; Shutt *et al.* 2017), as well as chikungunya (Perkins *et al.* 2015; Escobar *et al.* 2016; Del Valle *et al.* 2018), have drawn 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 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 (Perkins 2017; Asher *et al.* 2017) and anticipating future epidemics (Ferguson *et al.* 2016).

ACKNOWLEDGEMENTS

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).

LITERATURE CITED

Asher, J., Barker, C., Chen, G., Cummings, D., Chinazzi, M., Daniel-Wayman, S., *et al.* (2017). Preliminary results of models to predict areas in the Americas with increased likelihood of Zika virus transmission in 2017. *bioRxiv*.

Bjørnstad, O.N. & Grenfell, B.T. (2001). Noisy clockwork: time series analysis of population fluctuations in animals. *Science*, 293, 638–643.

Boletín Epidemiológico - Todos los documentos (2018). Available at: http://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/Forms/AllItems.aspx. Last accessed 18 February 2018.

Chretien, J.-P., Rivers, C.M. & Johansson, M.A. (2016). Make Data Sharing Routine to Prepare for Public Health Emergencies. *PLoS Med.*, 13, e1002109.

Cori, A. (2013). EpiEstim: a package to estimate time varying reproduction numbers from epidemic curves. R package version 1.1-2.

Cori, A., Ferguson, N.M., Fraser, C. & Cauchemez, S. (2013). A new framework and software to estimate time-varying reproduction numbers during epidemics. *Am. J. Epi.*, 178, 1505–1512.

Cummings, D.A.T., Irizarry, R.A., Huang, N.E., Endy, T.P., Nisalak, A., Ungchusak, K., *et al.* (2004). Travelling waves in the occurrence of dengue haemorrhagic fever in Thailand. *Nature*, 427, 344–347.

Dalziel, B.D., Pourbohloul, B. & Ellner, S.P. (2013). Human mobility patterns predict divergent epidemic dynamics among cities. *Proc. Biol. Sci.*, 280, 20130763.

Del Valle, S.Y., B.H. McMahon, J. Asher, R. Hatchett, J.C. Lega, H.E. Brown, *et al.* (2018) Summary results of the 2014-2015 DARPA chikungunya challenge. *BMC Infec. Dis.*, 18, 245.

Durrett, R. & Levin, S. (1994). The Importance of Being Discrete (and Spatial). *Theor. Popul. Biol.*, 46, 363–394.

Escobar, L.E., Qiao, H. & Peterson, A.T. (2016). Forecasting Chikungunya spread in the Americas via data-driven empirical approaches. *Parasit. Vectors*, 9, 112.

Ferguson, N.M., Cucunubá, Z.M., Dorigatti, I., Nedjati-Gilani, G.L., Donnelly, C.A., Basáñez, M.-G., *et al.* (2016). Countering the Zika epidemic in Latin America. *Science*, 353, 353–354.

Grenfell, B.T., Bjørnstad, O.N. & Kappey, J. (2001). Travelling waves and spatial hierarchies in measles epidemics. *Nature*, 414, 716–723.

Hastings, A. (2010). Timescales, dynamics, and ecological understanding. *Ecology*, 91, 3471–3480.

Huber, J.H., Childs, M.L., Caldwell, J.M. & Mordecai, E.A. (2017). Seasonal temperature variation influences climate suitability for dengue, chikungunya, and Zika transmission. *bioRxiv*, doi:10.1101/230383.

Instituto Nacional de Salud (INS). (2017). *Boletín Epidemiológico*. Available at: http://www.ins.gov.co/boletin-epidemiologico/Paginas/default.aspx. Last accessed 6 May 2017.

Keeling, M.J., Mezić, I., Hendry, R.J., McGlade, J. & Rand, D.A. (1997). Characteristic length scales of spatial models in ecology via fluctuation analysis. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 352, 1589–1601.

King, A.A., Domenech de Cellès, M., Magpantay, F.M.G. & Rohani, P. (2015). Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to Ebola. *Proc. Biol. Sci.*, 282, 20150347.

Koelle, K. & Pascual, M. (2004). Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera. *Am. Nat.*, 163,

901–913.
Kucharski, A.J., Funk, S., Eggo, R.M., Mallet, H.-P., Edmunds, W.J. & Nilles, E.J. (2016). Transmission Dynamics of Zika Virus in Island Populations: A Modelling Analysis of the 2013-14 French Polynesia Outbreak. *PLoS Negl. Trop. Dis.*, 10, e0004726.
Lessler, J., Chaisson, L.H., Kucirka, L.M., Bi, Q., Grantz, K., Salje, H., *et al.* (2016). Assessing the global threat from Zika virus. *Science*, 353, aaf8160.
Levin, S.A. (1992). The Problem of Pattern and Scale in Ecology. *Ecology*, 73, 1943–1967.
Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K. (2017). cluster: Cluster Analysis Basics and Extensions. R package version 2.0.6.
Metcalf, C.J.E., K.S. Walter, A. Wesolowski, C.O. Buckee, E. Shevliakova, A.J. Tatem, *et al.* (2017). Identifying climate drivers of infectious disease dynamics: recent advances and challenges ahead. *Proc. Roy. Soc. B*, 284, 20170901.
Mills, H.L. & Riley, S. (2014). The spatial resolution of epidemic peaks. *PLoS Comput. Biol.*, 10, e1003561.
Pascual, M. & Levin, S.A. (1999). From individuals to population densities: searching for the intermediate scale of nontrivial determinism. *Ecology*, 80, 2225–2236.
Perkins, T.A. (2017). Retracing Zika's footsteps across the Americas with computational modeling. *Proc. Natl. Acad. Sci. U. S. A.*, 114, 5558–5560.
Perkins, T.A., Metcalf, C.J.E., Grenfell, B.T. & Tatem, A.J. (2015). Estimating drivers of autochthonous transmission of chikungunya virus in its invasion of the Americas. *PLoS Curr.*
Perkins, T.A., Siraj, A.S., Ruktanonchai, C.W., Kraemer, M.U.G., Tatem, A.J. (2016). Model-based projections of Zika virus infections in childbearing women in the Americas. *Nat. Microbiol.*, 1, 16126.
Rand, D.A. & Wilson, H.B. (1995). Using spatio-temporal chaos and intermediate-scale determinism to quantify spatially extended ecosystems. *Proc. Roy. Soc. B: Biol. Sci.*, 259, 111–117.
Reynolds, A.P., Richards, G., de la Iglesia, B. & Rayward-Smith, V.J. (2006). Clustering Rules: A Comparison of Partitioning and Hierarchical Clustering Algorithms. *J. Math. Model. Algorithms*, 5, 475–504.
Rodriguez-Morales, A.J., García-Loaiza, C.J., Galindo-Marquez, M.L., Sabogal-Roman, J.A., Marin-Loaiza, S., Lozada-Riascos, C.O., *et al.* (2016). 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.*, 14, 286–288.
Rojas, D.P., Dean, N.E., Yang, Y., Kenah, E., Quintero, J., Tomasi, S., *et al.* (2016). The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. *Euro Surveill.*, 21.
Rousseeuw, P.J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.*, 20, 53–65.
Salje, H., Lessler, J., Maljkovic Berry, I., Melendrez, M.C., Endy, T., Kalayanarooj, S., *et al.* (2017). Dengue diversity across spatial and temporal scales: Local structure and the effect of host population size. *Science*, 355, 1302–1306.
Salje, H., Lessler, J., Paul, K.K., Azman, A.S., Rahman, M.W., Rahman, M., *et al.* (2016). 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.*, 113, 13420–13425.
Shutt, D.P., Manore, C.A., Pankavich, S., Porter, A.T. & Del Valle, S.Y. (2017). Estimating the

reproductive number, total outbreak size, and reporting rates for Zika epidemics in South and Central America. *Epidemics*, 21, 63–79.

Siraj, A.S., Rodriguez-Barraquer, I., Barker, C.M., Tejedor-Garavito, N., Harding, D., Lorton, C., *et al.* (2018). Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia. *Sci. Data*, 5, 180073.

Sistema de información geográfica para la planeación y el ordenamiento territorial (SIGOT). (2018). Available at: http://sigotn.igac.gov.co/sigotn/frames_pagina.aspx. Accessed Feb 2018.

Sorichetta, A., Hornby, G.M., Stevens, F.R., Gaughan, A.E., Linard, C. & Tatem, A.J. (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Sci. Data*, 2, 150045.

Turchin, P. & Taylor, A.D. (1992). Complex Dynamics in Ecological Time Series. *Ecology*, 73, 289–305.

Appendix S1. Description of simulation model of Zika virus transmission in Colombia.

We simulated data sets comparable to the observed data using an R implementation of the ZIKV transmission model described by Ferguson *et al.* (2016) parameterized to match the municipal-level R_0 values derived from Perkins *et al*. (2016). 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.* (2016).

To drive the model, we based estimates of the basic reproduction number, 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 (Perkins *et al.* 2016). 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 raster layer of human population projections (Sorichetta *et al.* 2015) aggregated to that scale by Siraj *et al*. (2018). We calibrated these R_0 estimates to observed dynamics in Colombia by scaling municipal values of R_0 from Perkins *et al.* (2016) by a constant (2.72) such that the value for the municipality of Girardot, Colombia, matched an estimate of 4.61 derived from an analysis of temporal incidence patterns there (Rojas *et al.* 2016). The environmental variables that drove spatial variation in these R_0 values include temperature, *Ae. aegypti* occurrence probability, and the gross cell product economic index.

To apply this model to Colombia, we used municipal-level human population sizes derived from WorldPop (Sorichetta *et al.* 2015) 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 by Ferguson *et al.* (2016) that we made was to remove explicit spatial coupling, given the complexity of doing so realistically for all 1,123 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). Although this approach was not able to capture differences in the timing of importation patterns across municipalities, none of the six features of the cumulative incidence curves that we analyzed depended on the timing of the epidemic in one municipality relative to another. To approximately match the national total of 85,353 suspected Zika cases, the time-varying ZIKV importation function that we used was scaled by a value of 1.55×10^{-3}. This value was obtained by trial-and-error tuning of example simulations in which a reporting rate of 11.5% was assumed (Kucharski *et al.* 2016). Also, given that our interest was in short-term dynamics rather than long-term dynamics as in Ferguson *et al.* (2016), we removed human age stratification from the model.

SUPPORTING TABLES

	Two groups		Three groups		Four groups	
	F_{836}	p	F_{836}	p	F_{836}	p
Ae. aegypti	19.7	9.9×10^{-6}	8.7×10^{-3}	0.93	0.86	0.35
$NDVI_{terra}$	13.3	2.8×10^{-4}	1.1	0.30	0.046	0.83
$NDVI_{aqua}$	12.9	3.5×10^{-4}	1.1	0.29	0.072	0.79
Mean temp.	32.9	1.4×10^{-8}	0.14	0.71	1.7	0.19
Pct. urban	29.6	7.0×10^{-8}	7.9	5.2×10^{-3}	0.029	0.86
Population	10.2	1.4×10^{-3}	1.7	0.19	0.50	0.48
GCP	8.3	4.1×10^{-3}	2.0	0.16	1.6	0.21
R_0	16.1	6.6×10^{-5}	0.056	0.81	1.7	0.19

Table S1. Summary of results from one-way analyses of variance at the municipal scale (n=836). For each relevant environmental variable (rows), we performed an analysis of variance to test for differences in the mean of that variable across two, three, or four groups identified by the classification analysis (columns). The F statistic and p value of each test is shown.

	Two groups		Three groups		Four groups	
	F_{31}	p	F_{31}	p	F_{31}	p
Ae. aegypti	2.9	0.10	0.24	0.63	0.38	0.54
$NDVI_{terra}$	1.2	0.29	0.022	0.88	$2.8x10^{-4}$	0.99
$NDVI_{aqua}$	0.52	0.48	0.32	0.58	0.21	0.65
Mean temp.	5.5	0.025	4.0	0.055	3.9	0.058
Pct. urban	0.28	0.60	0.37	0.54	0.42	0.52
Population	0.13	0.72	0.04	0.84	$8.0x10^{-3}$	0.93
GCP	1.5	0.23	1.5	0.24	1.1	0.31
R_0	3.1	0.086	0.38	0.54	1.0	0.32

Table S2. Summary of results from one-way analyses of variance at the departmental scale (n=31). For each relevant environmental variable (rows), we performed an analysis of variance to test for differences in the mean of that variable across two, three, or four groups identified by the classification analysis (columns). The F statistic and p value of each test is shown.

SUPPORTING FIGURES

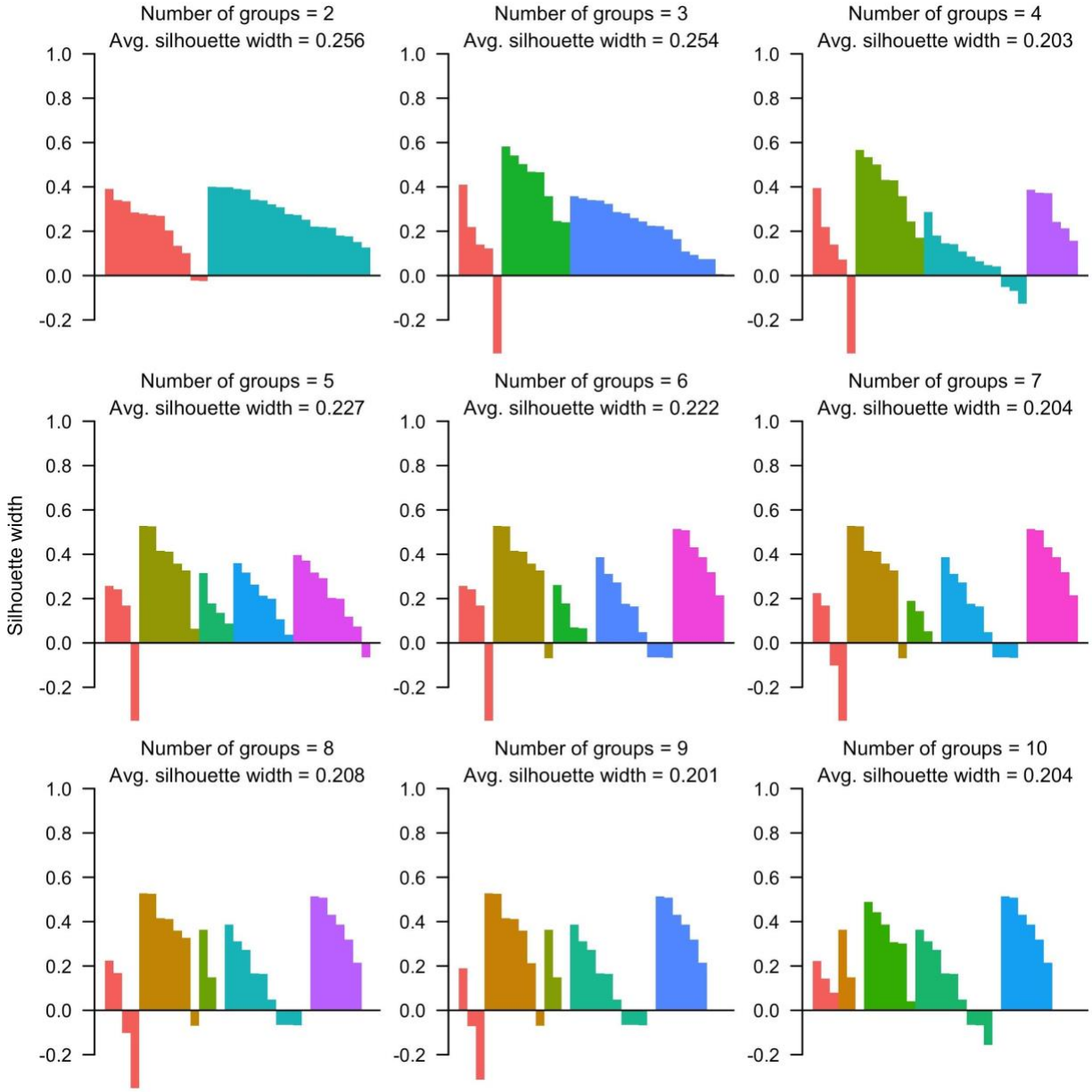

Figure S1. Silhouette plots at the departmental level for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given department according to the group assignments indicated by different colors in each panel. Higher average silhouette values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

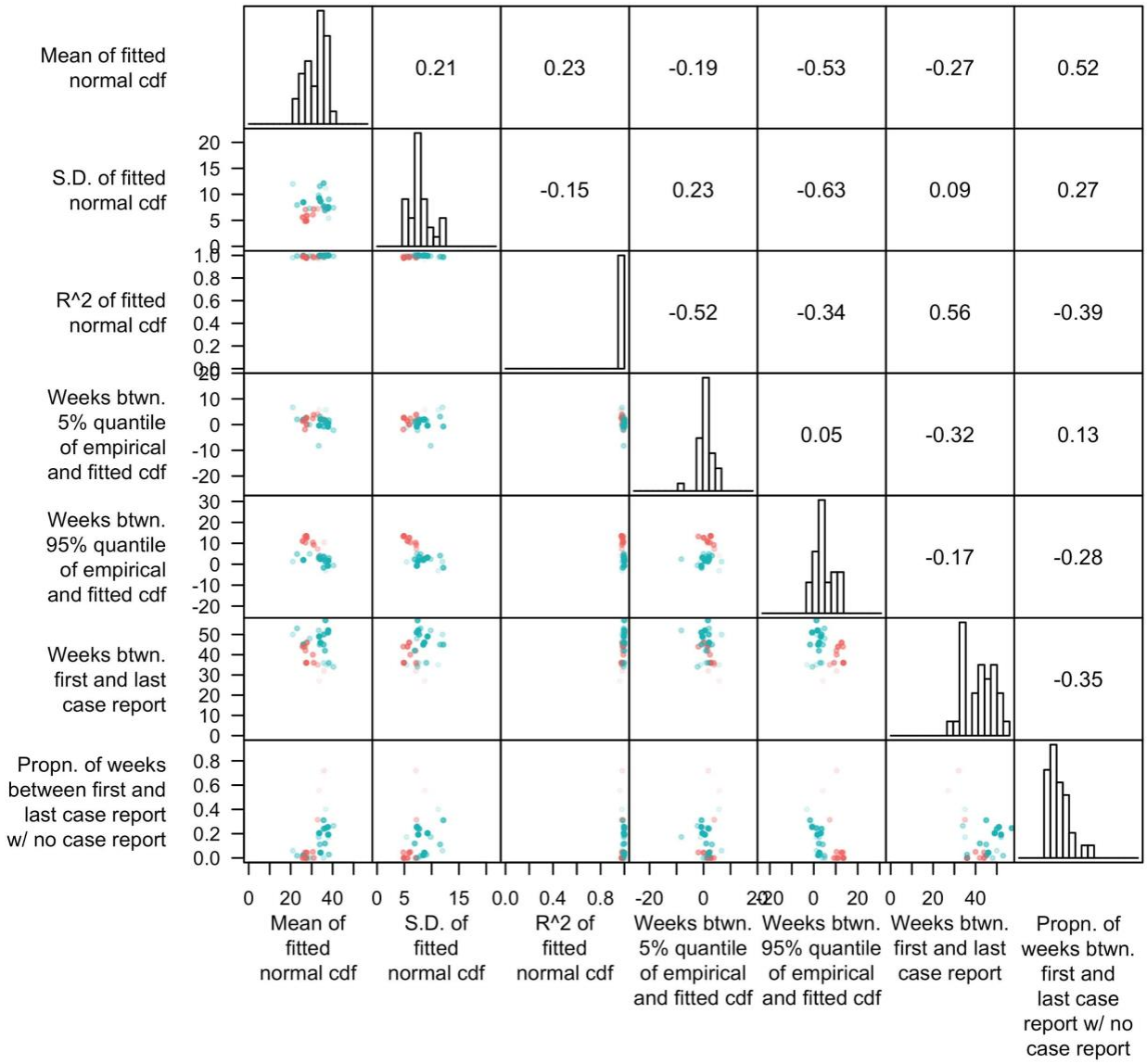

Figure S2. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the departments into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients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

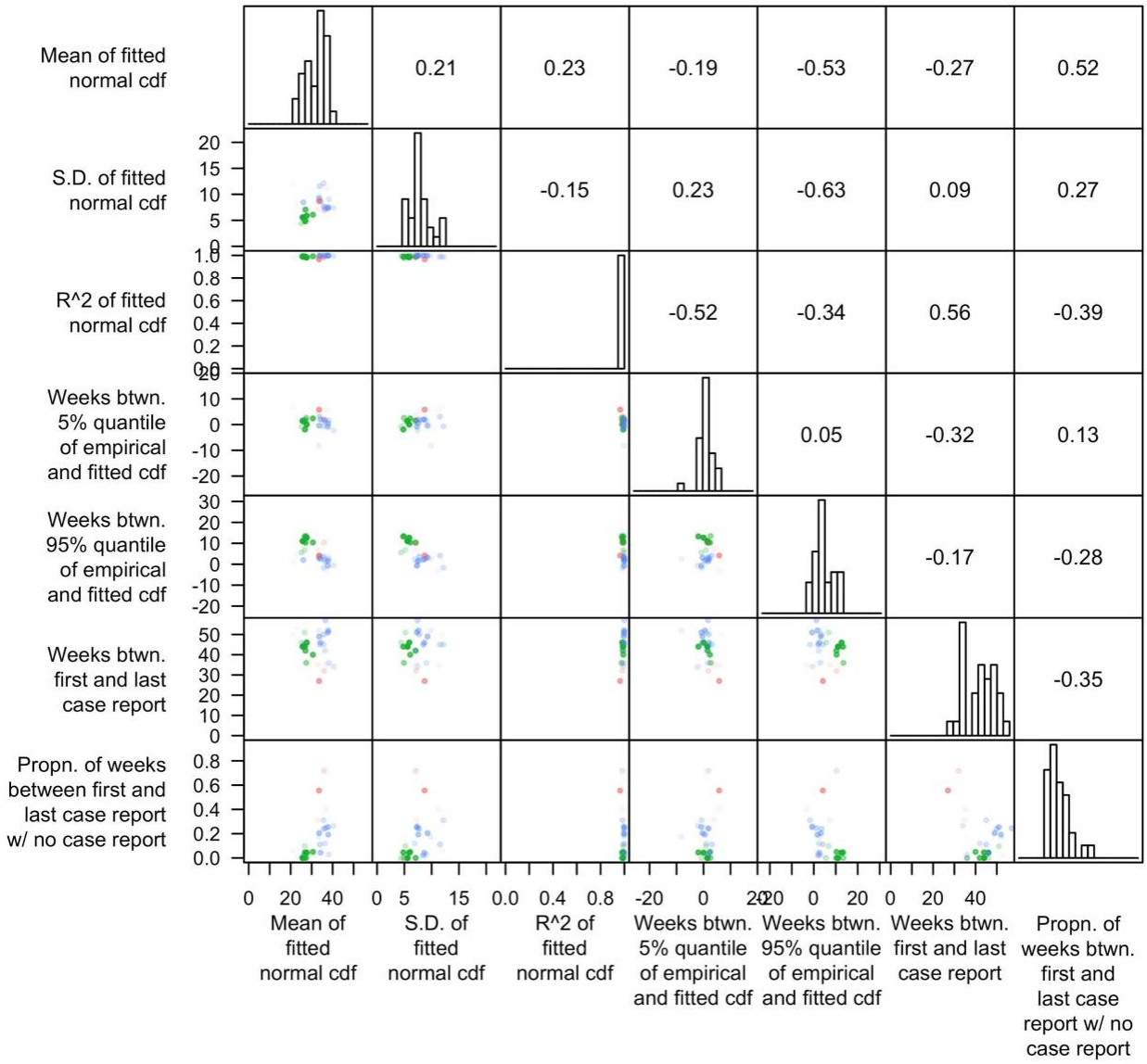

Figure S3. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the departments into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients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

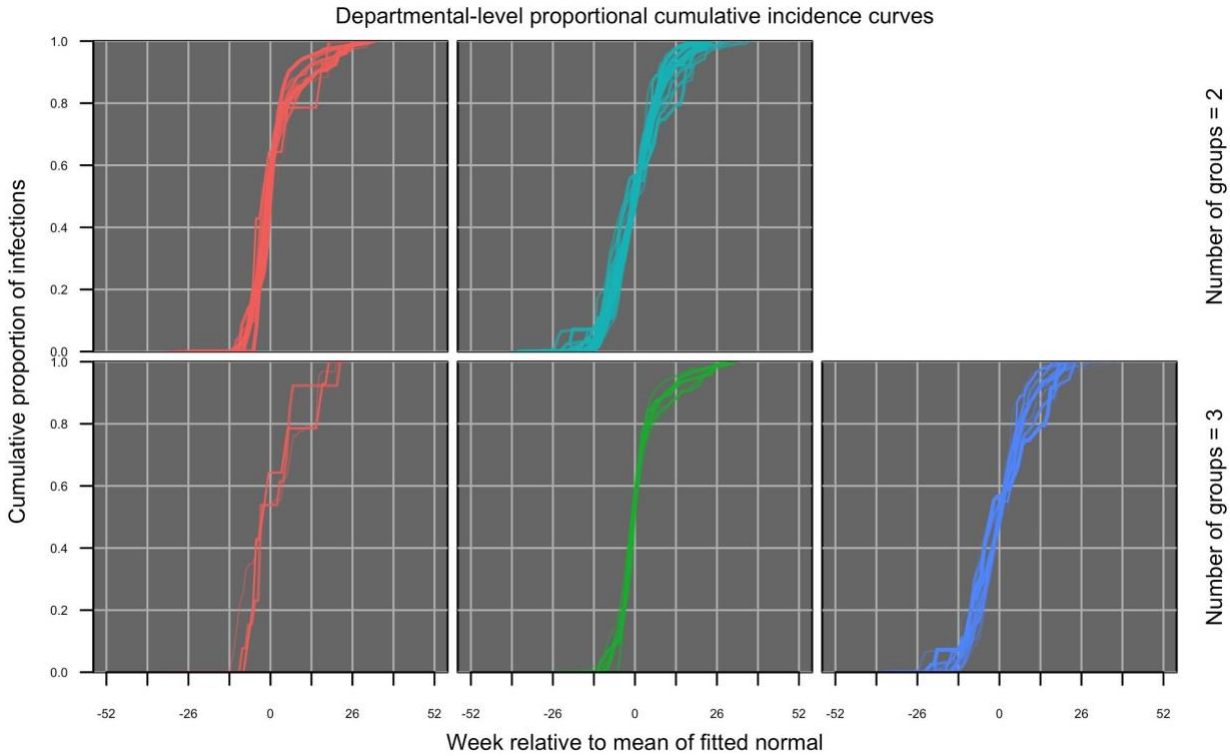

Figure S4. Proportional cumulative incidence curves at the departmental level with two (top) or three (bottom) groups. Within each row, groups are distinguished by color.

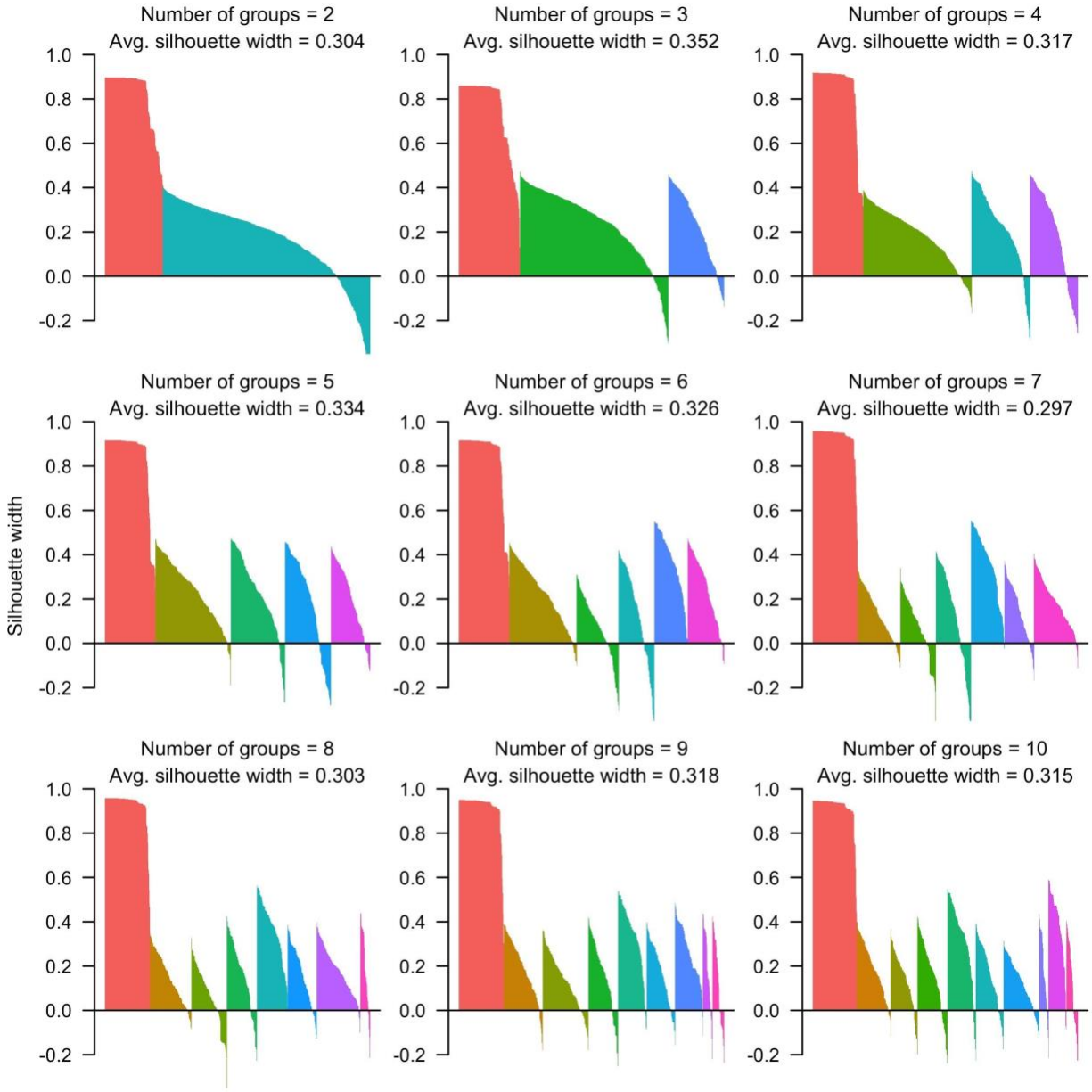

Figure S5. Silhouette plots at the municipal level for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given municipality according to the group assignments indicated by different colors in each panel. Higher average silhouette values indicate stronger clustering.

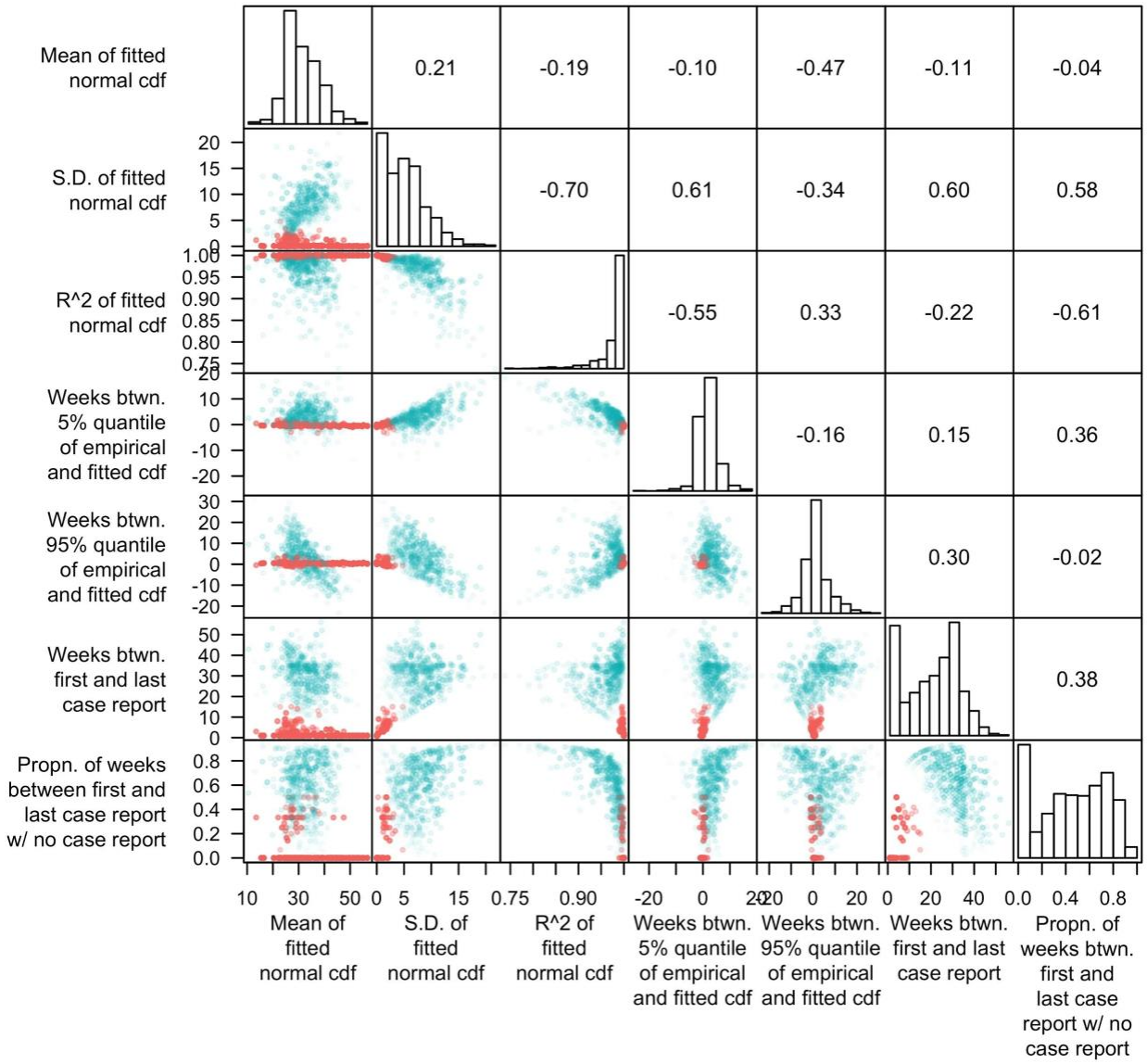

Figure S6. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients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

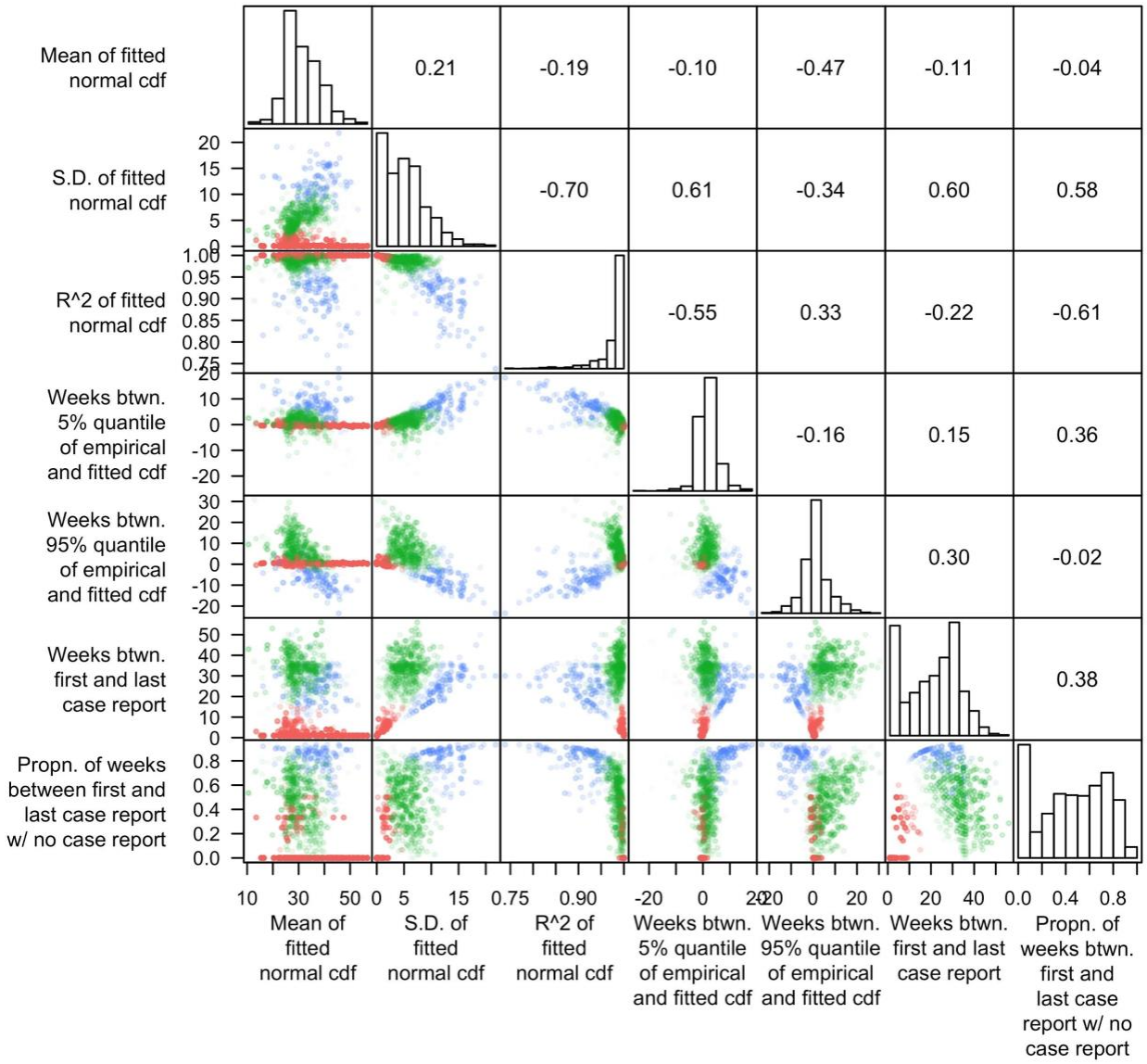

Figure S7. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation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

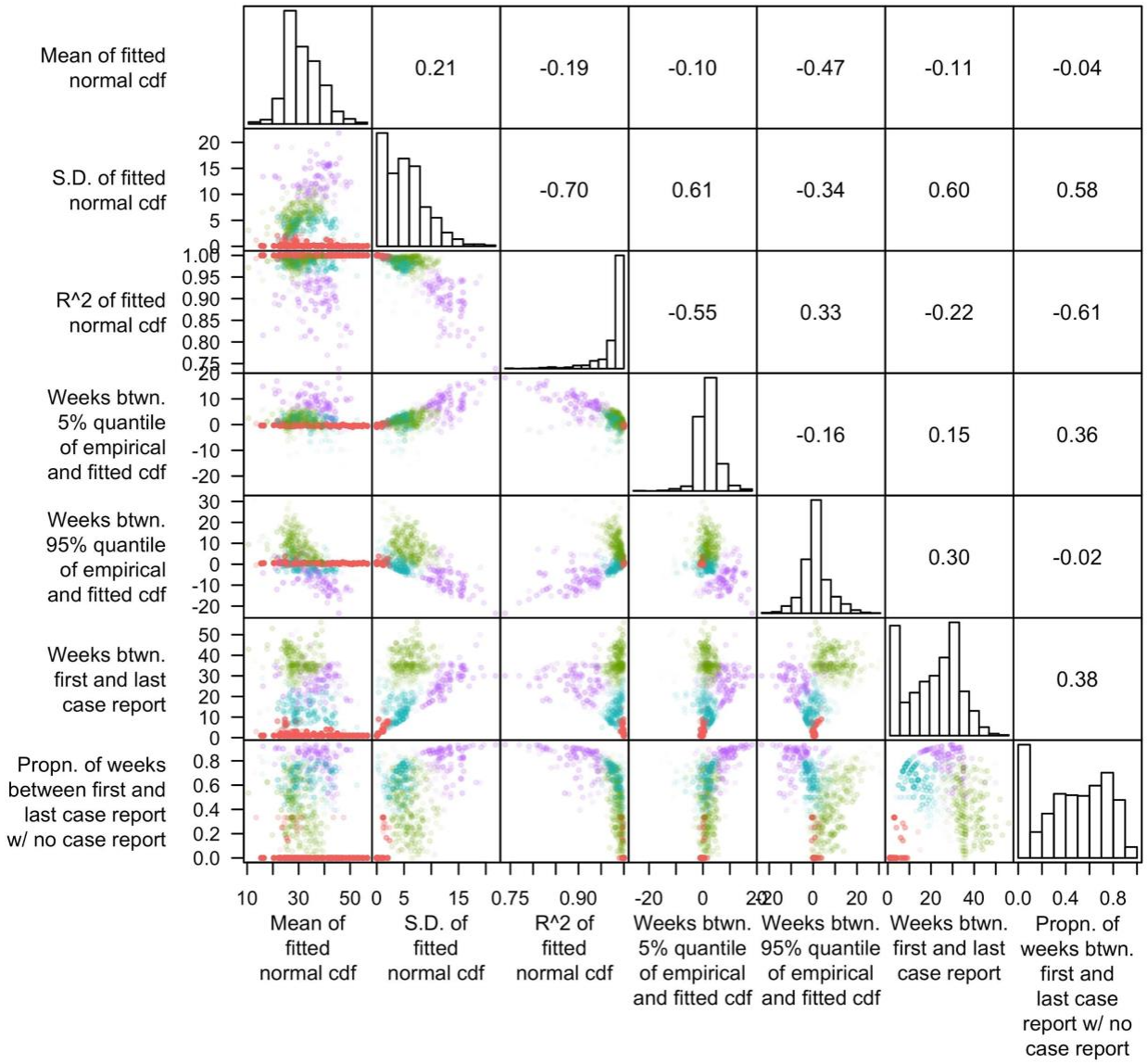

Figure S8. Pairwise plots of features of proportional cumulative incidence curves, with colors distinguishing group assignment of the municipalities into one of four groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

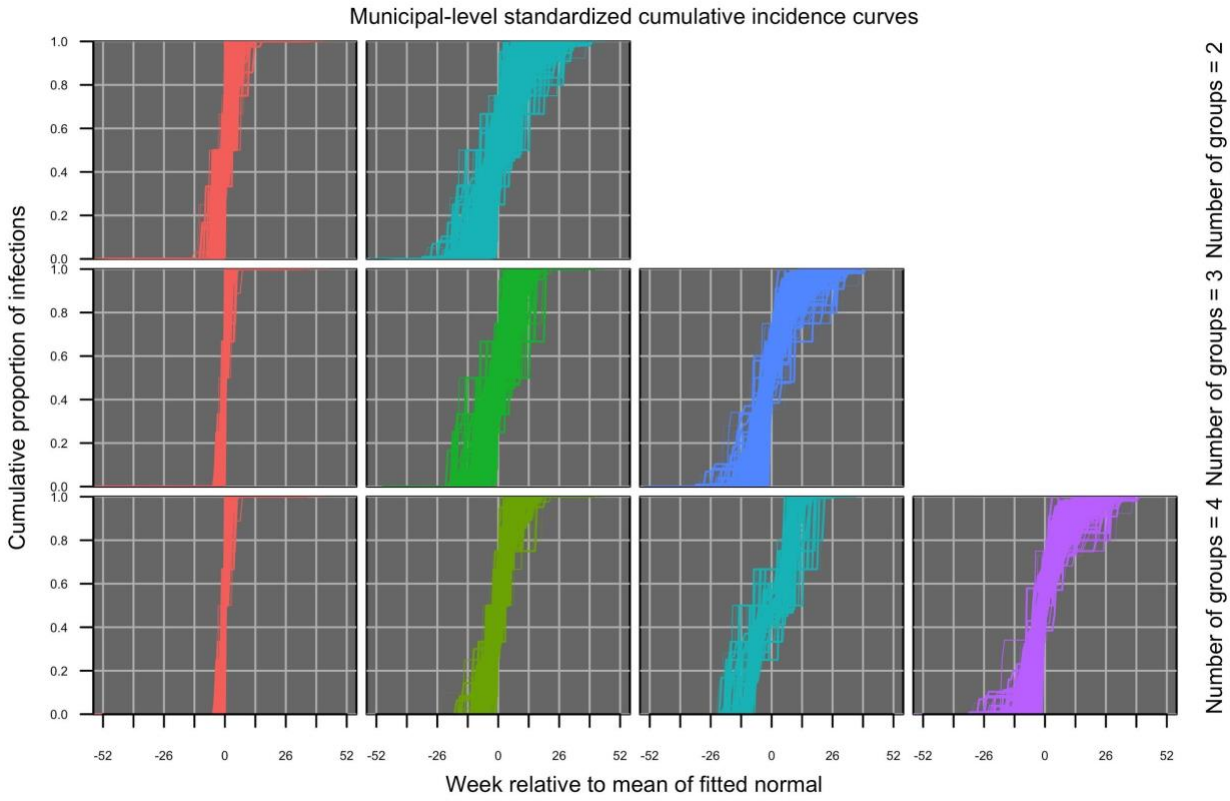

Figure S9. Proportional cumulative incidence curves at the municipal level with two (top), three (middle), or four (bottom) groups. Within each row, groups are distinguished by color.

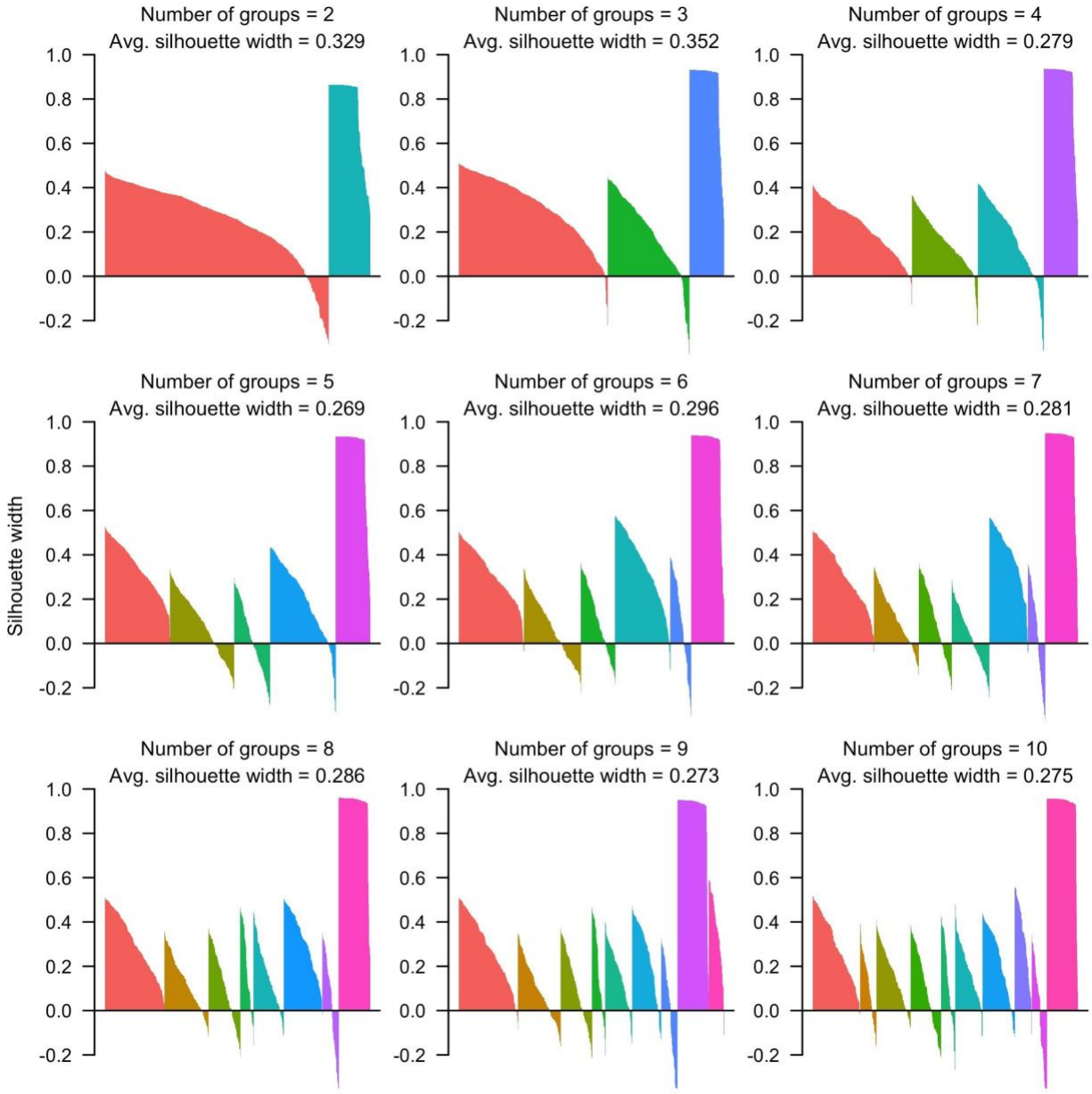

Figure S10. Silhouette plots at the municipal level based on a randomly selected simulated data set for groups numbering two to ten obtained by partitioning around medoids. Each bar corresponds to the silhouette value of a given municipality according to the group assignments indicated by different colors in each panel. Higher average silhouette values indicate stronger clustering.

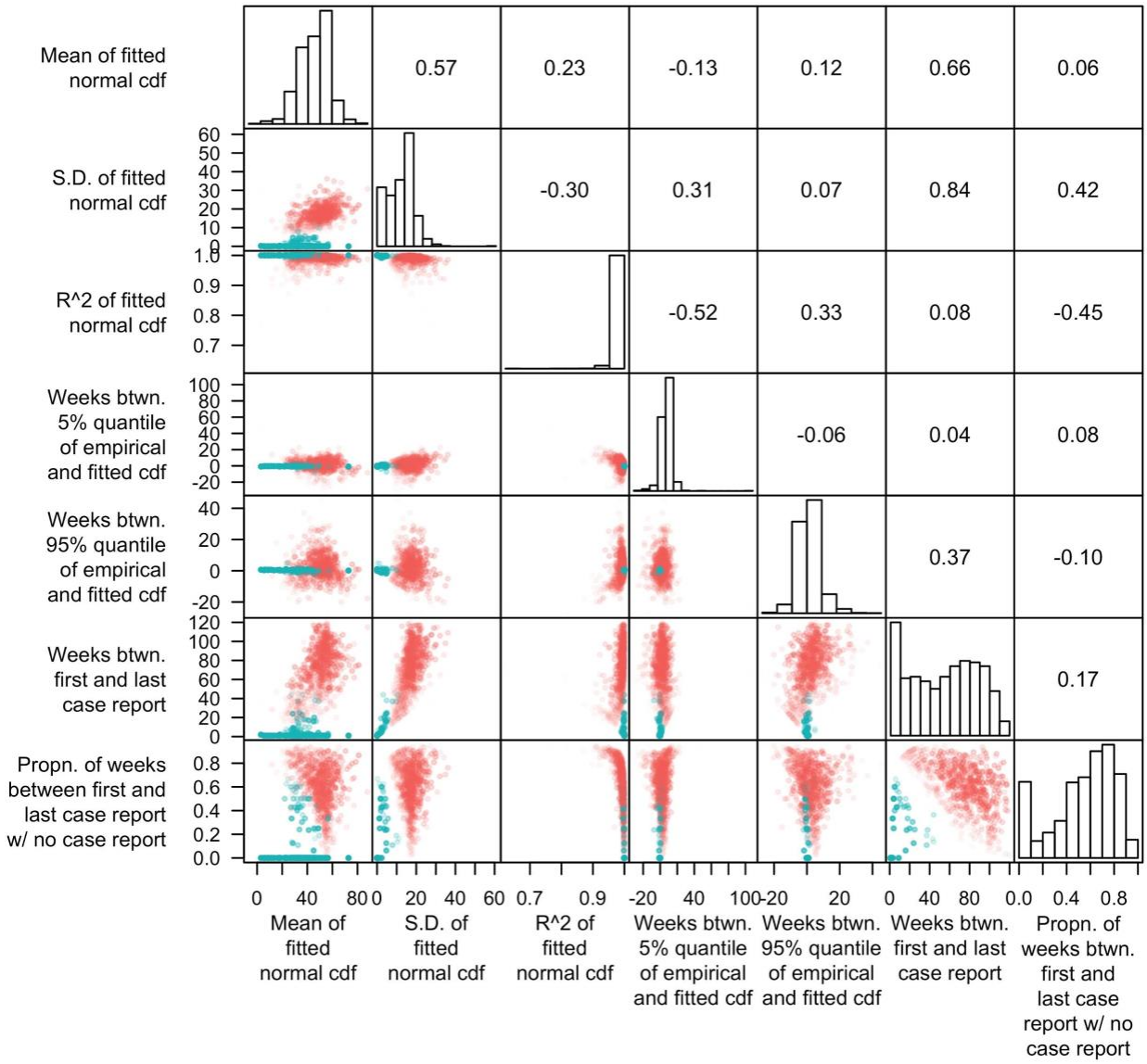

Figure S11. Pairwise plots of features of proportional cumulative incidence curves based on a randomly selected simulated data set, with colors distinguishing group assignment of the municipalities into one of two groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.

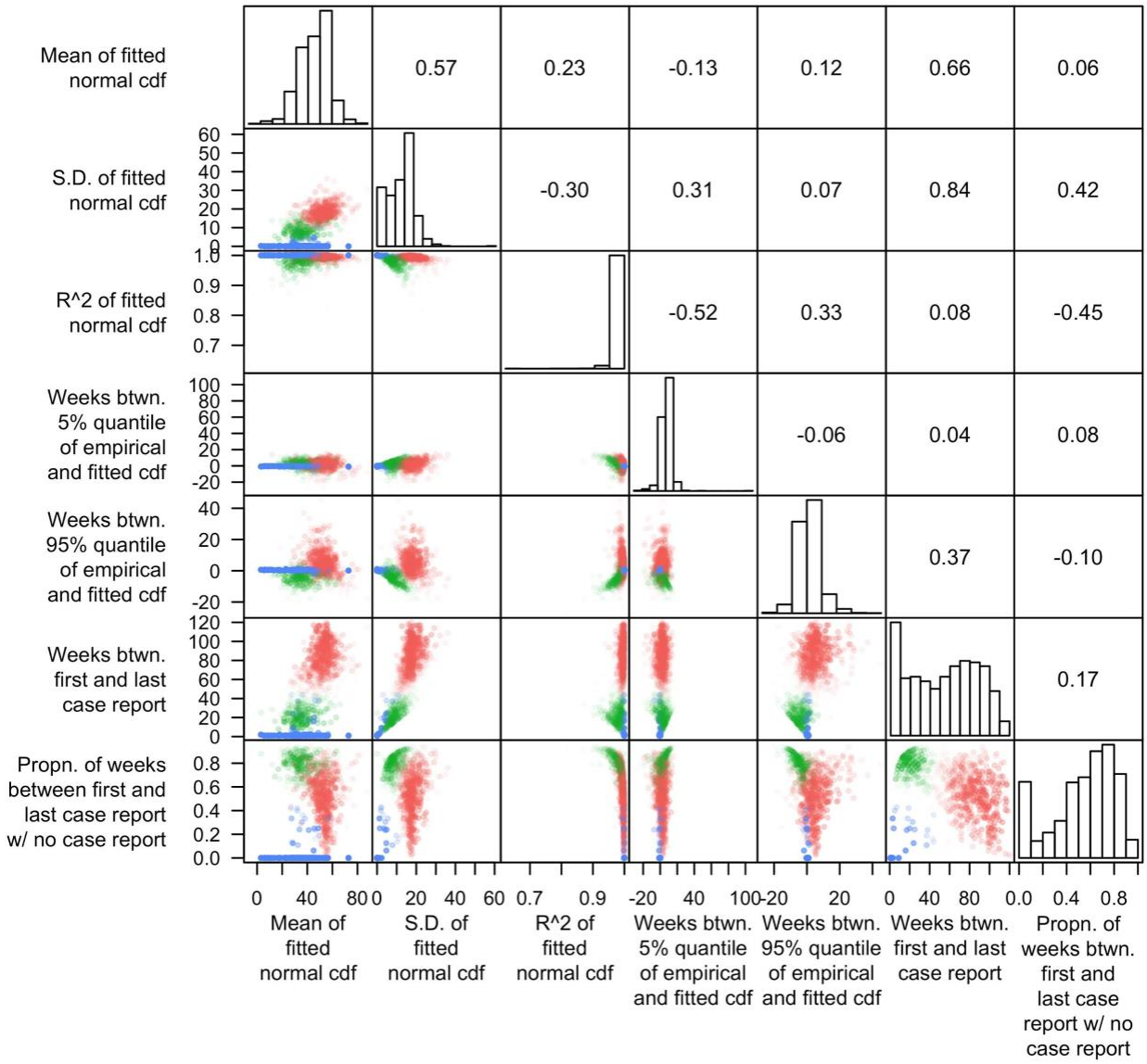

Figure S12. Pairwise plots of features of proportional cumulative incidence curves based on a randomly selected simulated data set, with colors distinguishing group assignment of the municipalities into one of three groups. Histograms show the marginal distributions of the features, and numbers in the upper right half indicate pairwise correlation coefficients between each pair of features. The transparency of each point is inversely proportional to silhouette value.