Kraemer et al. *EPJ Data Science* (2018) 7:16
https://doi.org/10.1140/epjds/s13688-018-0144-x

EPJ Data Science
a SpringerOpen Journal

EPJ.org

REGULAR ARTICLE Open Access

Inferences about spatiotemporal variation in dengue virus transmission are sensitive to assumptions about human mobility: a case study using geolocated tweets from Lahore, Pakistan

Moritz U.G. Kraemer[1,2,3*], D. Bisanzio[4,5], R.C. Reiner[6], R. Zakar[7], J.B. Hawkins[1,2], C.C. Freifeld[2,8], D.L. Smith[6,9], S.I. Hay[6], J.S. Brownstein[1,2] and T. Alex Perkins[10*]

*Correspondence:
kramer.moritz@gmail.com;
taperkins@nd.edu
[1]Department of Pediatrics, Harvard Medical School, Boston, USA
[10]Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, Notre Dame, USA
Full list of author information is available at the end of the article

Abstract

Billions of users of mobile phones, social media platforms, and other technologies generate an increasingly large volume of data that has the potential to be leveraged towards solving public health challenges. These and other big data resources tend to be most successful in epidemiological applications when utilized within an appropriate conceptual framework. Here, we demonstrate the importance of assumptions about host mobility in a framework for dynamic modeling of infectious disease spread among districts within a large urban area. Our analysis focused on spatial and temporal variation in the transmission of dengue virus (DENV) during a series of large seasonal epidemics in Lahore, Pakistan during 2011–2014. Similar to many directly transmitted diseases, DENV transmission occurs primarily where people spend time during daytime hours, given that DENV is transmitted by a day-biting mosquito. We inferred spatiotemporal variation in DENV transmission under five different assumptions about mobility patterns among ten districts of Lahore: no movement among districts, movement following patterns of geo-located tweets, movement proportional to district population size, and movement following the commonly used gravity and radiation models. Overall, we found that inferences about spatiotemporal variation in DENV transmission were highly sensitive to this range of assumptions about intra-urban human mobility patterns, although the three assumptions that allowed for a modest degree of intra-urban mobility all performed similarly in key respects. Differing inferences about transmission patterns based on our analysis are significant from an epidemiological perspective, as they have different implications for where control efforts should be targeted and whether conditions for transmission became more or less favorable over time.

Keywords: Big data; Disease dynamics; Geo-located tweets; Gravity model; Human mobility; Radiation model; Spatiotemporal analysis; Twitter data

1 Introduction

The spread and transmission dynamics of human infectious diseases are shaped extensively by human behavior [18]. Pathogen transmission depends on human contact patterns and tends to accelerate in highly connected areas with high population size and frequent travel [23]. Relevant population interactions between areas can be the result of daily commuting to the commercial center of a city and back [32, 50], visiting relatives or friends [27], religious or cultural activities [15], or many other reasons. Generally, urban travel is characterized by extensive daily activity, as work activities do not typically take place at the same places where people live [19]. Dynamic human movement patterns in cities can be inferred using a variety of data sources such as census data, mobile phone data, or social media data [28]. Passive data collection from social media platforms now offer timely, high resolution estimates of spatiotemporal patterns of human mobility [4, 5, 28, 32]. All of these movement types have the potential to shape infectious disease transmission dynamics, potentially in different ways depending on the mode of transmission (e.g., by direct contact or through a mosquito vector).

Specifically in the context of urban transmission, the importance of spatial heterogeneity in drivers of transmission is well-documented [40, 47, 54]. Some districts of a city may have considerably higher likelihood of infection as a result of, for example, higher mosquito densities (e.g., malaria [53], dengue [56]), yet each district contributes to transmission in any other district, not just within its own boundaries, due largely to human travel [11, 31, 34]. Such considerations can become highly relevant when human mobility is high, as observed in large urban areas [12] and can critically inform how resources to control and eliminate disease should be allocated [9, 11]. Understanding this interaction between human mobility, spatial variation in drivers of transmission, and control measures is important to know where control measures will be most impactful, as dengue [48], chikungunya [14, 49], yellow fever [21, 22], and Zika [30] continue to cause large urban outbreaks to control their spread and limit the burden caused by these viruses. An important feature that they have in common is that they are all transmitted by the *Aedes* mosquitoes, which are active during daytime hours when human mobility is high [54].

In this study, we examined a series of seasonal dengue epidemics in an urban setting that occurred between 2011 and 2014 in Lahore, Pakistan; no major epidemic had been recorded before that date [24]. Dengue virus (DENV) is a flavivirus transmitted between humans primarily by the *Aedes aegypti* mosquito [51]. Dengue burden is enormous and it has increased substantially in recent decades [6]. The distribution of *Ae. aegypti* is now larger than it has ever known to be [25], and the viruses it transmits have been expanding too as a result [29, 33], leading to expanding ranges or changes in the epidemiology of Zika, chikungunya, and yellow fever.

To enhance our understanding of urban transmission dynamics of infectious diseases and to evaluate the importance of assumptions of the spatial configuration of cities, we here use human mobility models and estimates derived from the social network platform Twitter to compare inferences about spatiotemporal variation in transmission patterns and determine how sensitive these inferences are to different assumptions about patterns of intra-urban human mobility. Little sensitivity of these inferences would suggest that analyses could proceed with business as usual assumptions, whereas strong sensitivity would point to a need for more careful consideration of human mobility data within analyses of infectious disease dynamics, even at the granularity of intra-urban scales.

2 Material & methods

Epidemiological data: We obtained individual dengue case data from Lahore, Pakistan, and aggregated to the town level (an administrative subdivision of the city, $n = 10$) on a weekly basis from January 1, 2011 to December 31, 2014. We refer to the number of dengue cases reported in town i in week t as $I_{i,t}$. Data were provided by the Health Department, Pakistan, and were processed from their original line list form. In total, 35,348 confirmed and suspected cases were recorded in Lahore. Roughly 18,020 of those occurred during the 2011 epidemic alone. Details of the geo-positioning procedure are described in detail in Kraemer et al. [26].

Human mobility data and models: To quantify human mobility patterns, we used openly available data from Twitter through its API. Our database consists of tweets made in Lahore from January 1, 2011 through June 30, 2015. Specifically, the tweets were gathered by querying the free streaming API for a bounding box of [−180, 180] longitude and [−90, 90] latitude, so all tweets with geographic coordinates match. The results are limited by Twitter to 1% of total tweet volume. We then filtered the database to only include tweets sent within the city of Lahore, Pakistan. The penetration of Twitter users with geo-located information amounts to about 1% of the total population in the study period, similar to previous estimates [28]. Other information included the user's unique ID. We associated each user with a town of residence according to which town they sent most tweets from during night hours defined as 9pm–7am.

To use the tweets to summarize mobility patterns of residents of the 10 different towns, we computed a single matrix H that contained the proportion of tweets made in town j by residents of town i, where i and j refer to the row and column of H, respectively. Thus, the rows of H sum to 1, and the columns of H sum to values somewhat less than or greater than 1. Due to the somewhat limited number of tweets available from users in a given town during a given time period and because there was no obvious seasonality in the data, we did not make use of temporally disaggregated Twitter data in our transmission model.

In addition to the H matrix based on tweets, we constructed four alternative H matrices that span a wide range of assumptions about human mobility commonly used in infectious disease modeling. At one extreme, we constructed an H matrix following the ideal free assumption that movements between all locations occur proportional to population size. At the opposite extreme, we constructed an H matrix consistent with an assumption of no movement between towns. Just as the H matrix based on tweets represents an intermediate assumption between these two extremes, we formulated two additional H matrices based on commonly used models of human mobility; the gravity model [60] and the radiation model [50]. We applied these models to data about the distance between town centroids and town population sizes. This produced values of fluxes between i and j but did not produce an estimate of the magnitude of time spent in i by residents of i. To work around this gap in the predictions of these models, we used the diagonal of the tweet-based H matrix as the diagonal for these two H matrices. For the off-diagonal elements, we normalized the fluxes out of i predicted by the gravity and radiation models and multiplied those terms by $1 - H_{i,i}$. Numerical values of all five H matrices are provided in Tables S1–S5 (Additional file 1).

Mobility-based transformation of incidence data: Our analysis is premised on the distinction between the location where an individual resides and the locations where she or he spends time. DENV is transmitted by the urban-adapted mosquito *Ae. aegypti*, which

engages in the majority of its blood feeding activity during daytime hours [1]. Because this means that transmission is expected to occur mainly where people spend time during the day rather than where they reside [55], we transformed the ten residence-based incidence time series $I_{i,t}$ to ten mobility-based incidence time series $\tilde{I}_{i,t}$. The latter contains the incidence of cases acquired in town i in week t under a given assumption about mobility patterns defined by H and is calculated according to $\tilde{I}_{i,t} = \sum_j H_{j,i} I_{j,t}$. We examined a total of five different interpretations of $\tilde{I}_{i,t}$ corresponding to the five different assumptions about human mobility patterns quantified by five different H matrices, as described in the previous section.

Transmission model: We used a spatial TSIR framework to model the dynamics of $\tilde{I}_{i,t}$ in the ten towns of Lahore during 2011–2014. Consistent with the assumptions about mobility used to define $\tilde{I}_{i,t}$, we defined the effective population size of town i during daytime hours as $\tilde{N}_i = \sum_j H_{j,i} N_j$. We are not aware of any significant DENV transmission activity in Lahore prior to 2011, so we assumed that the effective number of susceptible individuals in town i during daytime hours was $\tilde{S}_{i,1} = \tilde{N}_i$ during the first week of January, 2011. Thenceforth, the susceptible population was depleted as new cases arose according to $\tilde{S}_{i,t} = \tilde{S}_{i,t-1} - \tilde{I}_{i,t}/\rho$, where ρ is the probability that a person infected with DENV reported to the Health Department. Although there is a great deal of variability in ρ due to variation in rates of symptomatic disease and health-seeking behavior in different populations, we adopted a value of $\rho = 0.18$ based on a recent meta-analysis [10]. This parameter accounts for the fact that many DENV infections are mild or asymptomatic, which is important when tracking the susceptible population due to the fact that individuals exposed to DENV become immune thereafter regardless of the extent to which they experience symptoms. One complication that we did not account for due to a lack of data is that there are four distinct DENV serotypes, with long-lasting immunity being specific only to the serotype(s) to which one has been exposed. There is, however, a short-term period of cross-immunity that is protective against all serotypes following exposure to only a single serotype, with the duration of this period (maximum-likelihood estimate: 1.88 y, 95% confidence interval: 0.88–4.31 y [43]) being similar to the timescale of our data set as a whole.

Following the standard form of TSIR models, we assumed that new cases among people spending time in town i were acquired on week t according to

$$\tilde{I}_{i,t} = \beta_i(t) \frac{\tilde{I}'_{i,t}}{\tilde{N}_i} \tilde{S}'_{i,t}, \tag{1}$$

where $\beta_i(t)$ is the transmission coefficient in town i at time t. The prime notation for $\tilde{I}'_{i,t}$ and $\tilde{S}'_{i,t}$ denotes the numbers of infected and susceptible people in the "generation" prior to t. The obligatory role of a mosquito in the transmission of DENV from one person to another is associated with a relatively long generation interval compared to directly transmitted pathogens. Whereas most TSIR models treat consecutive time steps as distinct generations, we obviated the need to temporally aggregate the data to such a large extent by calculating

$$\tilde{I}'_{i,t} = \sum_{n=1}^{5} \omega_n \tilde{I}_{i,t-n}, \tag{2}$$

where

$$\omega_n = \frac{1}{F(35)}\left(\frac{1}{7}\int_{7(n-1)}^{7n} F(\tau+7) - F(\tau)\,d\tau\right) \tag{3}$$

is the probability that a case in week t is attributable to a case that occurred in week $t-n$ as defined by a generation interval with distribution function F [38]. We adopted a distribution function estimated by Siraj et al. [52] at a temperature of 30°C (the average daily temperature in Lahore during 2011–2014), which resulted in values of ω_n of $4.8 \times 10^{-4}, 0.168, 0.440, 0.267$, and 0.125 for $n = 1, \ldots, 5$, respectively. $\tilde{S}'_{i,t}$ was calculated similarly.

Model fitting: A primary advantage of the TSIR framework is that it allows for a model to be fitted to incidence data using regression techniques, which are easier to implement than alternative approaches to fitting dynamic models to time series data. To do so, we took the natural log of Eq. (1) and rearranged to obtain the regression equation

$$\ln(\tilde{I}_{i,t}) - \ln\left(\tilde{S}'_{i,t}\right) + \ln(\tilde{N}_i) = \ln\left(\beta_i(t)\right) + \ln\left(\tilde{I}'_{i,t}\right) + \epsilon_t, \tag{4}$$

where

$$\ln\left(\beta_i(t)\right) = s_{\text{secular},i}(t) + s_{\text{seasonal},i}\left(t - 52\lfloor t/52\rfloor\right) \tag{5}$$

and each ϵ_t is an independent and identically distributed normal random variable. We posed $\ln(\beta_i(t))$ as a shape-constrained additive model (SCAM) and estimated parameters describing its two components using the scam function in the scam package [41] in R [42]. To prevent data points near the beginning and end of the time series from leading to unreasonably large values of $\beta_i(t)$ when extrapolating beyond those data points, we constrained $s_{\text{secular},i}$ to be a concave function. We modeled $s_{\text{seasonal},i}$ as a cyclic cubic spline to ensure that its values at the beginning and end of the year were equal up to their second derivative. Under all mobility assumptions other than ideal free, we estimated separate town-specific functions for each of the two components of $\ln(\beta_i(t))$. Under the ideal free mobility assumption, we estimated only a single $\ln(\beta(t))$ that applied to all towns due to the fact that the mobility-transformed data were strictly proportional to each other under this assumption.

3 Results

Human mobility data: Tweet-derived movement estimates showed relatively low movement outside the town of residence, with the mean proportion of time spent within one's town of residence being 91.2% (range: 84.0–96.8%). The town from which the largest proportion of non-resident tweets was made was Gulberg Town (1.7%), and the fewest were made in Wagha Town (0.16%). Although there was substantial day-to-day variation in Twitter activity across towns (Fig. 1), the extent to which that variation was driven by a set of deterministic factors or sampling noise was not apparent. Based on the limited sample of tweets available and their incomplete coverage over the study period, we used the time-averaged proportion of tweets by residents of each town made in every other town in the epidemiological analysis.

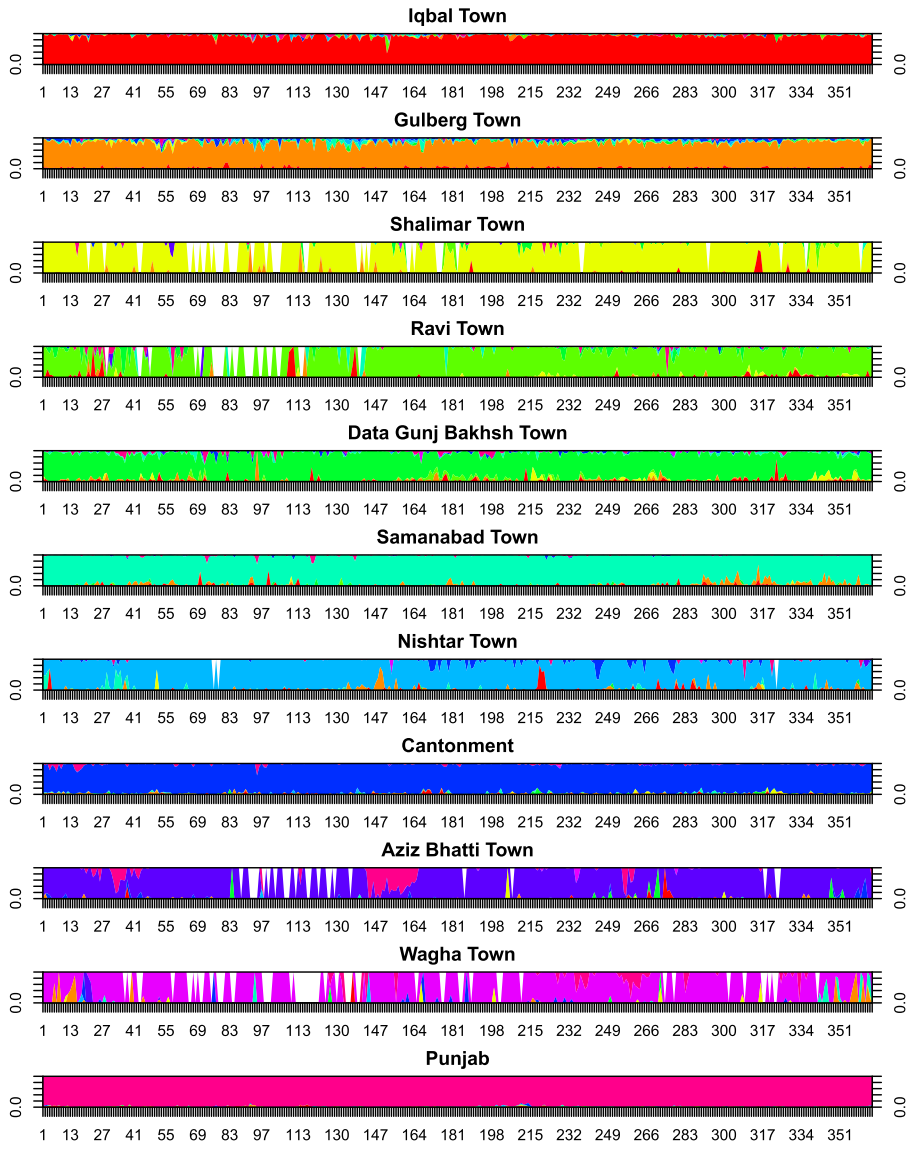

Figure 1 Relative proportion of tweets made in the town indicated in the panel label by residents of every other town. Different colors represent different home location of residents

Mobility-based transformation of incidence data: Applying the five mobility matrices to dengue incidence time series stratified by town of residence, we obtained notably different time series of the towns in which the cases were acquired. Under the assumption of no movement outside one's town of residence, the residence-based and mobility-based time series were identical. Under the assumption that mobility follows Twitter, gravity, or radiation movement patterns, the mobility-based time series was mostly similar to the residence-based time series (Fig. 2), although redistribution from high-incidence towns to low-incidence towns was visually apparent (Fig. 3). This redistribution was attributable to the partially homogenizing effect of inter-town mobility. Under the assumption that mobility follows an ideal-free distribution, the mobility-based time series was stratified proportional to town population size (Fig. 2), resulting in time series that followed identi-

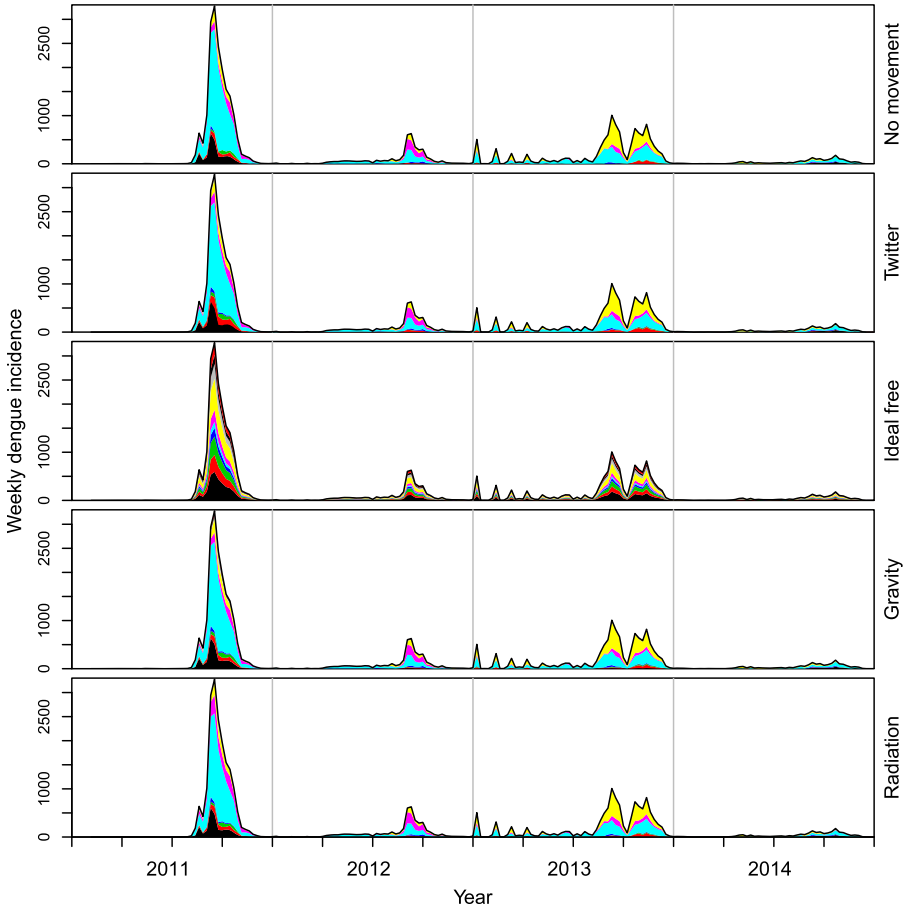

Figure 2 Time series stratified by towns in which cases were assumed to be acquired (colors) under five different assumptions about mobility among towns (rows)

cal dynamics (Fig. 3). Whereas the distribution of incidence across towns was temporally constant under the ideal-free assumption, there was substantially more temporal variation in the distribution of incidence across towns under the Twitter, gravity, radiation, and no-movement assumptions (Fig. 2).

Model fitting: Best-fit models to all five mobility-based time series explained a relatively high proportion of variation in incidence, with the coefficient of determination, R^2, ranging 0.519–0.685 (Fig. 4). In general, the time series data was explained similarly well by each of the models that performed a mobility transformation; i.e., Twitter ($R^2 = 0.678$), ideal free ($R^2 = 0.662$), gravity ($R^2 = 0.685$), and radiation ($R^2 = 0.662$). The data was explained less well by the model that assumed no movement ($R^2 = 0.519$). Rather than an indication of the inadequacy of the no movement assumption, we interpreted this lower R^2 value as a consequence of the fact that $\tilde{I}_{i,t}$ in Eq. (1) is integer-valued under the no movement assumption and continuous under the other mobility assumptions. Because our model's generation-interval adjustment in Eq. (2) results in $\tilde{I}'_{i,t}$ being continuous, the mobility assumptions associated with continuous values of $\tilde{I}_{i,t}$ have an inherent advantage in fitting the data, especially for $\tilde{I}'_{i,t} < 1$ (Fig. 4). As a result, comparison of R^2 values calculated in reference to the mobility-based time series to which the models were fitted

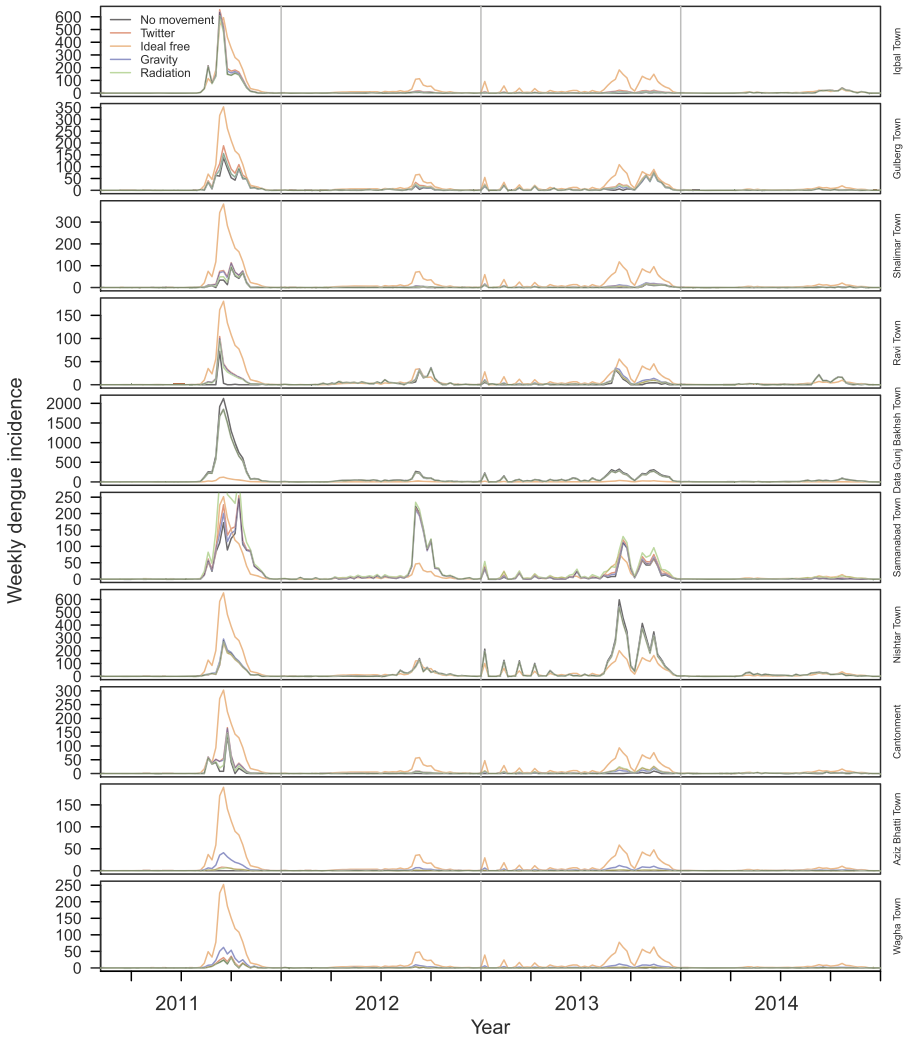

Figure 3 Time series stratified by towns in which cases were assumed to be acquired (rows) under five different assumptions about mobility among towns (line type)

indicates no clear distinction among the mobility assumptions and their appropriateness for modeling the mobility-based time series.

Another way that we examined model fit was based on how well each best-fit model matched the original time series once a model's one-step ahead predictions of $\tilde{I}_{i,t}$ were transformed back to predictions of $I_{j,t}$ using the H matrix under which a given model was fitted. Under the no movement assumption, $\tilde{I}_{i,t}$ and $I_{j,t}$ were, by definition, the same. These predictions were generally consistent with the data, although there were instances in Gulberg Town, Shalimar Town, and Wagha Town in which model predictions far exceeded the data during certain time periods (Fig. 5). This was likely due to the seasonal component of $\beta_i(t)$ being influenced too heavily by years with larger outbreaks. Under the Twitter, gravity, and radiation assumptions, predictions of $I_{j,t}$ were often similar to or nearly as good as predictions based on the model fitted under the no movement assumption (Fig. 5). These models performed less well in the towns with the lowest incidence—i.e., Aziz Bhatti Town and Wagha Town—due to those models' predictions of a greater degree of imported inci-

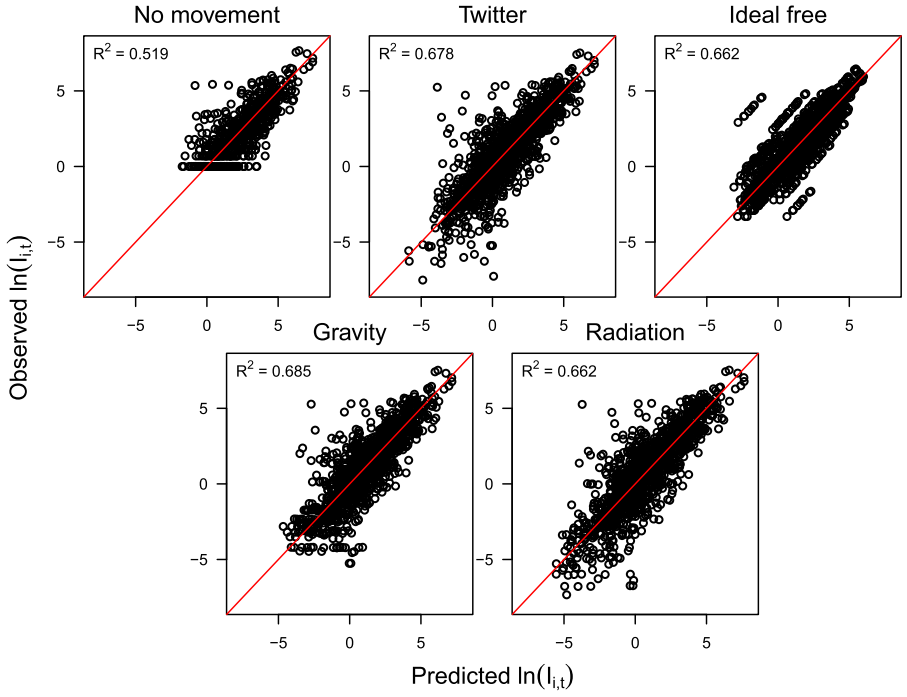

Figure 4 Relationships between predicted (*x*-axis) and observed (*y*-axis) log incidence based on models fitted to five different mobility-based incidence time series (panels). The coefficient of determination, R^2, associated with each best-fit model is indicated in each panel. Values of observed incidence vary across panels due to the effect of different assumptions about mobility used to transform the residence-based time series to mobility-based time series. For example, log incidence under the assumption of no movement never falls below 0, because there were no fractional cases observed in the raw data. Fractional incidence did occur in the other two time series due to each person's incidence of disease being partitioned across the towns proportional to assumed mobility patterns. Under the ideal free assumption, the diagonal sets of points are a result of incidence on a given day varying across towns only in proportion to their different population sizes

dence from towns with high transmission than actually occurred (Fig. 5). In terms of ability to predict $I_{j,t}$, the model fitted under the ideal free assumption performed the worst by far. In the towns with the greatest incidence per capita, this model predicted either too few cases overall (Samanabad Town) or incidence patterns that were not as peaked as was observed locally (Ravi Town) (Fig. 5). In all other towns, the model fitted under the ideal free assumption overpredicted incidence, sometimes by several hundred cases in a single week (e.g., Iqbal Town, Nishtar Town) (Fig. 5).

Transmission model inferences: Inferences of $\beta_i(t)$ under different mobility assumptions varied widely. Variation in $\beta_i(t)$ across towns was maximized under the assumption of no movement, with patterns ranging from nearly flat in Cantonment and Aziz Bhatti Town to a single seasonal peak in Shalimar Town and Wagha Town to multiple annual peaks of different heights across years in the other towns (Fig. 6, left). Under the no movement assumption, confidence intervals for $\beta_i(t)$ were unreasonably high in Shalimar Town and Wagha Town (Fig. 6, left). By design, inferences of $\beta_i(t)$ under the ideal free assumption were identical across towns and displayed a pattern of two seasonal peaks, with the one in the third quarter being larger (Fig. 6, right). This same general pattern was apparent in the inferences of $\beta_i(t)$ under the Twitter mobility assumption, but there was clear variability across towns, with differences in the heights of the peaks, their timing, and other

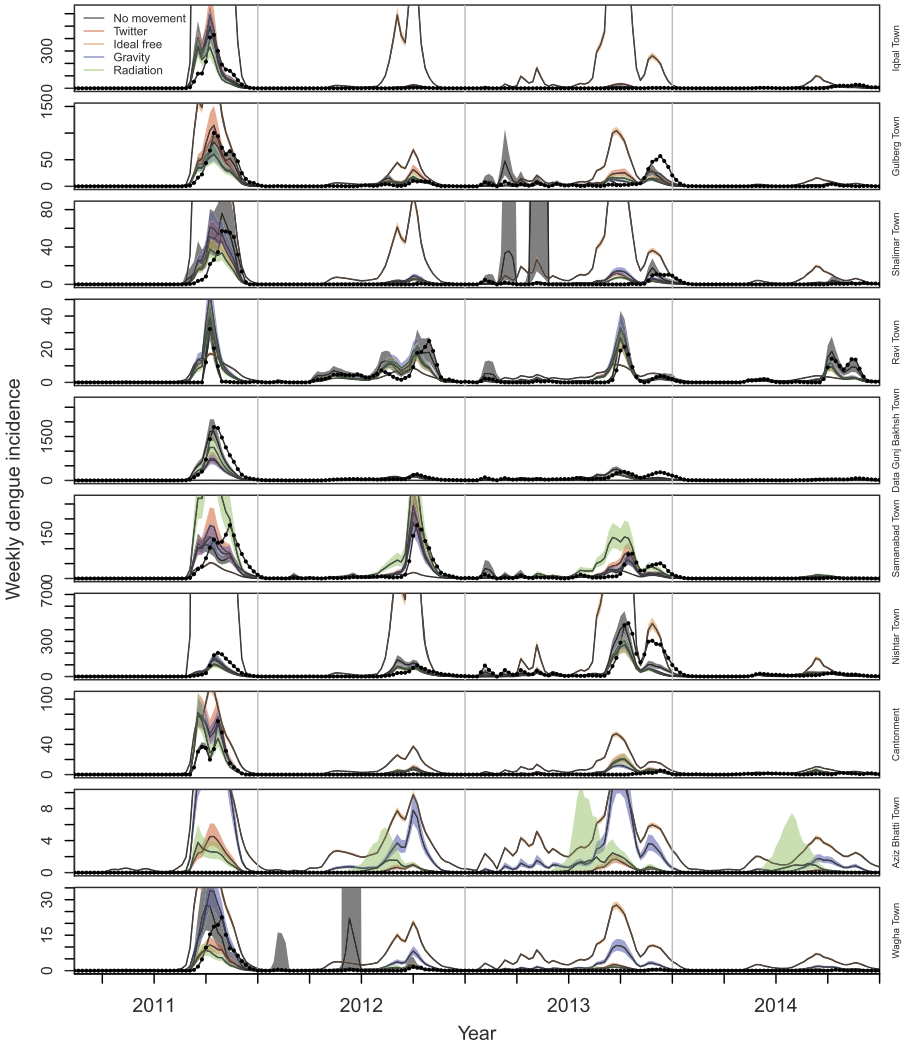

Figure 5 Predicted values of residence-based incidence in each town (rows) using the best-fit model under each of five different assumptions about mobility (colors). Observed values of residence-based incidence in each town are shown with black dots, and bands show 95% confidence intervals on model predictions

aspects of their shape (Fig. 6, center). Inferences of $\beta_i(t)$ under the gravity and radiation mobility assumptions were similar to those under the Twitter assumption, with gravity being extremely similar (Fig. 7, left) and radiation being similar in most towns but having considerably larger peaks in Aziz Bhatti Town and Wagha Town (Fig. 7, right).

A general tendency for variation in $\beta_i(t)$ across towns under different assumptions about mobility was reinforced by examining geometric means of $\beta_i(t)$ over time. Under the ideal free mobility assumption, the geometric mean of $\beta_i(t)$ decreased every year (Fig. 8). In contrast, the geometric mean of $\beta_i(t)$ was greater in 2014 than in 2011 in approximately half the towns under the Twitter, gravity, and radiation mobility assumption, with differences across models in terms of which towns experienced those increases (Fig. 8). The degree of inter-annual variation in the geometric mean of $\beta_i(t)$ was greatest under the no movement assumption, moderate under the Twitter, gravity, and radiation assumptions, and least under the ideal free assumption (Fig. 8). Under the Twitter, gravity, and radiation

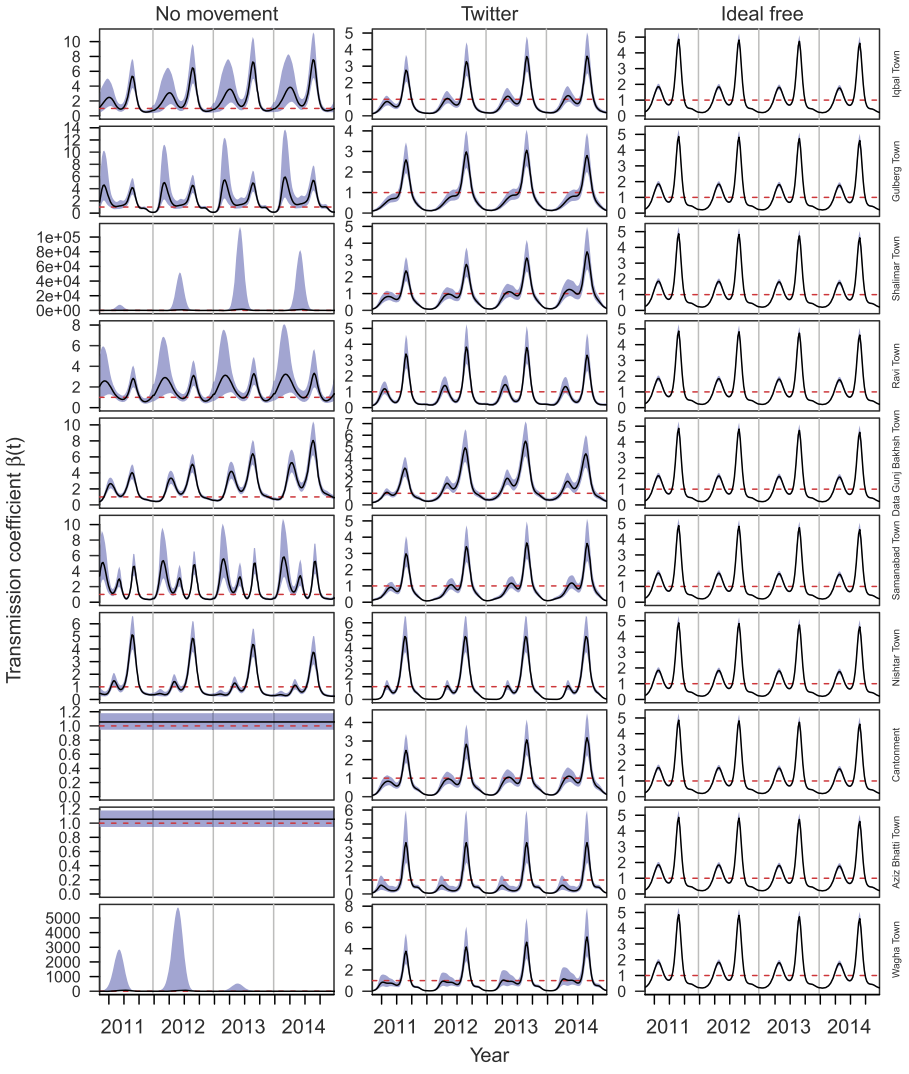

Figure 6 Temporal variation in $\beta_i(t)$ for different towns (rows) under three different mobility assumptions (columns): no movement (left), Twitter (center), and ideal free (right). The black line shows the mean of the best-fit model and blue bands show standard error around the mean. The dashed red line indicates where $\beta_i(t) = 1$

assumptions, the geometric mean of $\beta_i(t)$ across all years was highest in Data Gunj Bakhsh Town, which was also highest in both absolute and per capita terms for both residence-based and mobility-based incidence (Fig. 9). Otherwise, there was little correspondence between the geometric mean of $\beta_i(t)$ and either absolute or per capita incidence of the other nine towns under the Twitter, gravity, and radiation mobility assumptions. There was also no clear correspondence between $\beta_i(t)$ and incidence under the no movement assumption, and both the geometric mean of $\beta_i(t)$ and per capita incidence were equal across towns under the ideal free assumption, as expected (Fig. 9).

4 Discussion

Urban areas exhibit spatial heterogeneity in numerous factors that are relevant to infectious disease transmission, which can contribute to spatial variation in transmission [2]

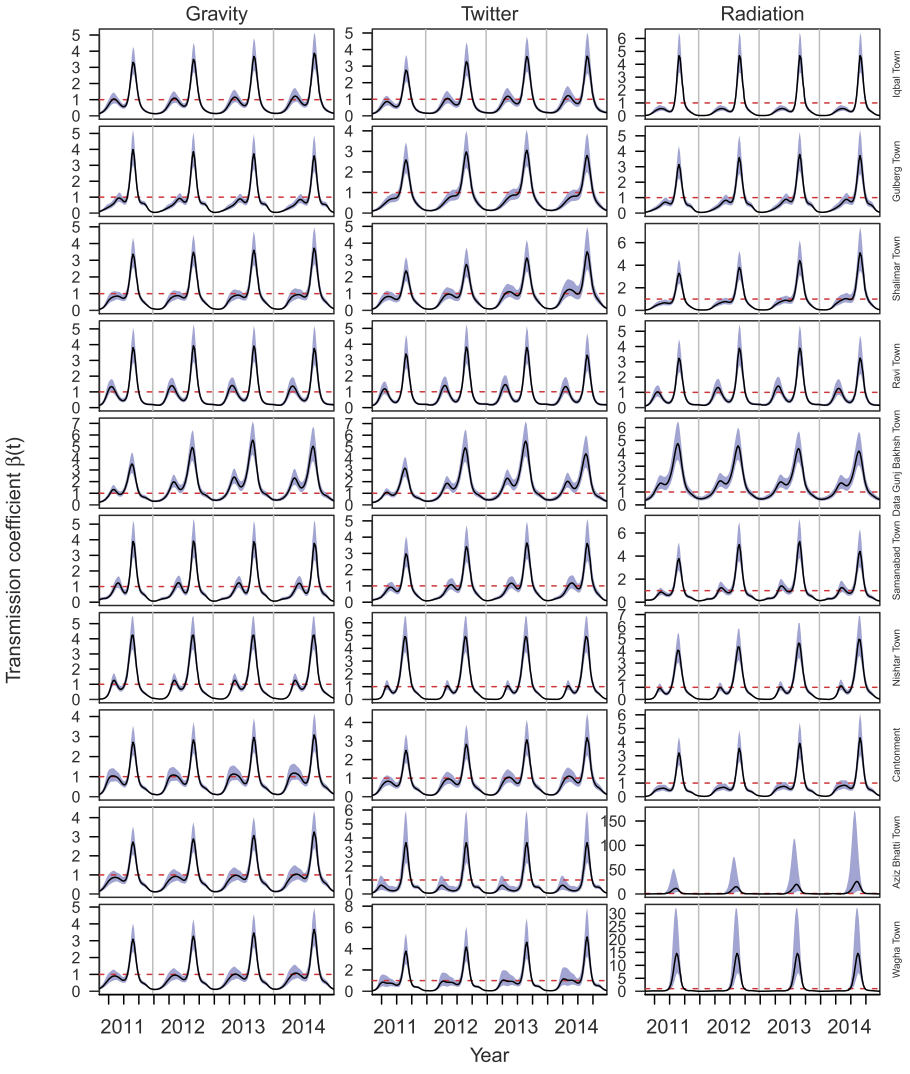

Figure 7 Temporal variation in $\beta_i(t)$ for different towns (rows) under three different mobility assumptions (columns): gravity model (left), Twitter (center), and radiation model (right). The black line shows the mean of the best-fit model and blue bands show standard error around the mean. The dashed red line indicates where $\beta_i(t) = 1$

and interact with temporal drivers of transmission [44]. Our work contributes to understanding of infectious disease dynamics in urban settings by highlighting the important role that human mobility plays in relating observed patterns of disease incidence to inferred patterns of disease transmission. On the one hand, our results show that assuming that human mobility is well mixed at the scale of a city (ideal free assumption) fails to capture underlying spatial heterogeneity in transmission and can lead to incorrect conclusions about secular trends in transmission across years. In some ways, this behavior is not surprising, but a systematic review of the literature on mathematical modeling of mosquito-borne disease transmission showed that this assumption is extremely prevalent across this field [45]. On the other hand, assuming that different districts of a city are isolated (no movement assumption) may lead to exaggerated and biologically unrealistic inferences about transmission patterns. Whether it be Twitter or other data streams [3],

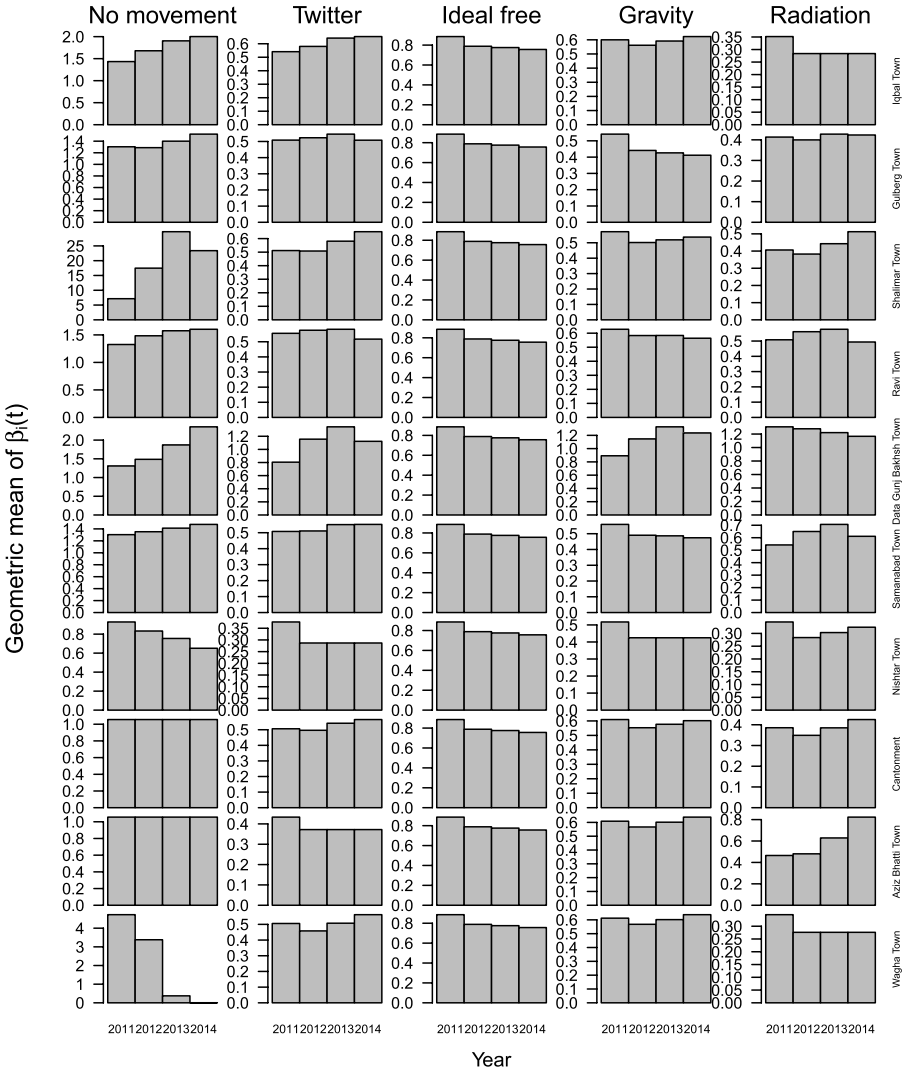

Figure 8 Geometric means of $\beta_i(t)$ stratified by year (x-axis), town (rows), and mobility assumption (columns)

incorporating realistic patterns of human mobility among districts of a city may help strike an appropriate balance between the tendencies of these two extreme assumptions. Regardless of whether such data streams provide a "correct" picture of mobility, it is encouraging that our results showed that Twitter, gravity, and radiation assumptions all resulted in similar epidemiological inferences.

The results of our analysis are a useful case study for any infectious disease, but they have particular importance for dengue. First, dengue mitigation strategies tend to be spatially reactive to reported incidence [59]. Other than the town with the highest per capita incidence, we did not identify a strong correspondence between towns with the highest incidence and those with the highest inferred transmission coefficients (similar to recent findings for malaria [11]), which suggests that reactive control deployed to areas with the highest incidence may not necessarily have as much impact on reducing transmission as more optimized strategies might. Second, due to adverse outcomes associated with vaccinating individuals with no prior DENV exposure with the only currently licensed dengue

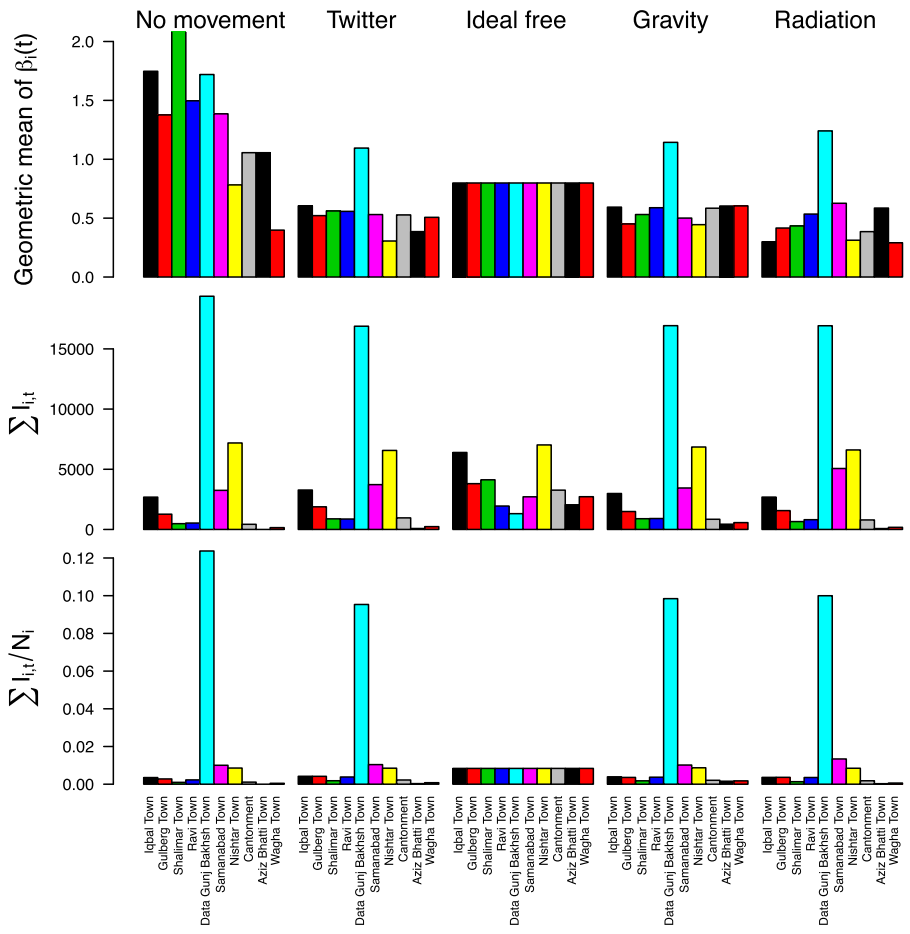

Figure 9 Measures of transmission and incidence across towns (colors) under different mobility assumptions (columns) aggregated over the entire 2011–2014 time period. The upper left y-axis was cut off to permit viewing of the majority of values; the geometric mean of $\beta_i(t)$ for Shalimar town is 17.49

vaccine [20], it is recommended that this vaccine only be used in areas with high transmission intensity [17]. Although consideration is given to subnational variation in transmission intensity (see https://mrcdata.dide.ic.ac.uk/_dengue/dengue.php), our results indicate that this issue may also warrant attention at intra-urban scales. Third, the increasing trend in the transmission coefficient that we observed under the Twitter mobility assumption serves as a reminder that a decreasing trend in incidence may not be indicative of a decreasing trend in factors underlying transmission. Instead, it appears that incidence has probably decreased due to an increase in herd immunity and that conditions remain ripe for transmission, which could result in a large epidemic once a sufficient number of susceptibles build up from births and waning heterotypic immunity.

Although our Twitter mobility assumption may be an improvement over some of the other assumptions that we examined, there are a number of other considerations about intra-urban human mobility that are likely to affect DENV transmission. We have limited understanding of how representative Twitter patterns are compared to actual movements of people and see them more as an approximation to relative movements between towns. At the same time, Twitter data have the advantage of being widely accessible for many urban areas worldwide, whereas alternative models for intra-urban human mobility

tend to have been developed around specific settings (e.g., [37]). In addition, there may be interactions between disease symptoms, infectiousness, and mobility in DENV-infected people [13, 39] that complicate the assumption that tweets by presumably healthy people are a suitable approximation of mobility patterns of people involved in transmission [58]. Higher order descriptions of movement may also be necessary to accurately capture transmission dynamics, as social network structure has been shown to affect transmission dynamics in urban environments [46, 54, 57]. There is also the perennial question of what spatial scale is satisfactory for modeling infectious disease dynamics [36, 40]. Examinations of intra-urban DENV transmission patterns in Bangkok, Thailand suggest that there can be strong spatial heterogeneities of relevance to transmission dynamics at scales as small as hundreds of meters [47, 49].

Overall, our results showed that the inferred degree of variation in a spatially and temporally variable transmission coefficient was sensitive to five different assumptions about intra-urban mobility that we considered. This approach extends previous applications of the TSIR model that estimated either seasonally varying transmission coefficients according to pre-defined functions (e.g., [16]) or completely independent values of the transmission coefficient at each time step (e.g., [35]) by blending the same underlying conceptual approach with a powerful new regression technique [41] and applying it in a spatial context. Although our analysis does not account for specifically which factors underlie variation in the transmission coefficient that we uncovered, there are many well-known candidates that could be incorporated into future analyses [7, 8, 25, 52]. Either way, we expect that our results about the sensitivity of transmission inferences to assumptions about intra-urban mobility would still apply. More generally, we hope that this case study will serve as a guiding example to the growing number of data scientists engaging in analyses of infectious disease dynamics. The increasing availability of data from Twitter and other Internet-based streams provide an exciting opportunity for extracting new understanding from time series of infectious disease incidence, if used within an appropriate conceptual framework.

Additional material

Additional file 1: Supplementary information (PDF 109 kB)

Acknowledgements
We are thankful to Dr. Irfan Ahmed, the World Health Organization (WHO), Punjab Office and Punjab Health Department for providing dengue case data.

Funding
MUGK is supported by The Branco Weiss Fellowship—Society in Science, administered by the ETH Zurich and acknowledges funding from a Training Grant from the National Institute of Child Health and Human Development (T32HD040128) and the National Library of Medicine of the National Institutes of Health (R01LM010812, R01LM011965). RCR, DLS, and TAP received support from a grant from the Bill and Melinda Gates Foundation (OPP 1110495 to DLS). TAP received support from a grant from the National Institutes of Health/National Institute of Allergy and Infectious Disease (1P01AI098670-01A1) and a DARPA Young Faculty Award (D16AP00114). JSB and JBH acknowledge support by the National Institutes of Health (NIH 5R01LMO011965-03).

Availability of data and materials
Data and code are available upon request.

Competing interests
The authors declare no competing interests.

Authors' contributions
MUGK, TAP designed the experiment. RCR, DLS advised on methods. SIH, JH, CCF, JSB, DB, RZ provided and processed data. MUGK, TAP performed the analyses and wrote the paper. All authors provided feedback and contributed to revisions. All authors read and approved the final manuscript.

Author details
[1]Department of Pediatrics, Harvard Medical School, Boston, USA. [2]Computational Epidemiology Lab, Boston Children's Hospital, Boston, USA. [3]Department of Zoology, University of Oxford, Oxford, UK. [4]RTI International, Washington, USA. [5]Center for Tropical Diseases, Sacro Cuore-Don Calabria Hospital, Negrar, Italy. [6]Institute for Health Metrics and Evaluation, University of Washington, Seattle, USA. [7]Department of Public Health, University of Punjab, Lahore, Pakistan. [8]College of Computer and Information Science, Northeastern University, Boston, USA. [9]Sanaria Institute for Global Health and Tropical Medicine, Rockville, USA. [10]Department of Biological Sciences and Eck Institute for Global Health, University of Notre Dame, Notre Dame, USA.

Publisher's Note

Received: 31 January 2018 Accepted: 31 May 2018 Published online: 11 June 2018

References

1. Akram W, Hafeez F, Ullah UN, Kim YK, Hussain A, Lee JJ (2009) Seasonal distribution and species composition of daytime biting mosquitoes. Entomol Res 39:107–113
2. Alexander L, Jiang S, Murga M, González MC (2015) Origin—destination trips by purpose and time of day inferred from mobile phone data. Transp Res, Part C, Emerg Technol 58:240–250
3. Althouse BM, Scarpino SV, Meyers LA, Ayers JW, Bargsten M, Baumbach J et al (2015) Enhancing disease surveillance with novel data streams: challenges and opportunities. EPJ Data Sci 4:17
4. Bassolas A, Lenormand M, Tugores A, Gonçalves B, Ramasco JJ (2016) Touristic site attractiveness seen through Twitter. EPJ Data Sci 5:12
5. Beiró MG, Panisson A, Tizzoni M, Cattuto C (2016) Predicting human mobility through the assimilation of social media traces into mobility models. EPJ Data Sci 5:30
6. Bhatt S, Gething PW, Brady OJ, Messina JP, Farlow AW, Moyes CL et al (2013) The global distribution and burden of dengue. Nature 496:504–507
7. Brady OJ, Golding N, Pigott DM, Kraemer MU, Messina JP, Reiner RC et al (2014) Global temperature constraints on *Aedes aegypti* and *Ae. albopictus* persistence and competence for dengue virus transmission. Parasites Vectors 7:338
8. Brady OJ, Johansson MA, Guerra CA, Bhatt S, Golding N, Pigott DM et al (2013) Modelling adult *Aedes aegypti* and *Aedes albopictus* survival at different temperatures in laboratory and field settings. Parasites Vectors 6:351
9. Chowell G, Nishiura H (2014) Transmission dynamics and control of Ebola virus disease (EVD): a review. BMC Med 12:196
10. Clapham HE, Cummings DAT, Johansson MA (2017) Immune status alters the probability of apparent illness due to dengue virus infection: evidence from a pooled analysis across multiple cohort and cluster studies. PLoS Negl Trop Dis 11:e0005926
11. Cohen JM, Le Menach A, Pothin E, Eisele TP, Gething PW, Eckhoff PA et al (2017) Mapping multiple components of malaria risk for improved targeting of elimination interventions. Malar J 16:459
12. Çolak S, Lima A, Gonzalez MC (2016) Understanding congested travel in urban areas. Nat Commun 7:10793
13. Duong V, Lambrechts L, Paul RE, Ly S, Lay RS, Long KC et al (2015) Asymptomatic humans transmit dengue virus to mosquitoes. Proc Natl Acad Sci USA 112:14688–14693
14. Faria NR, Lourenco J, Marques de Cerqueira E, Maia de Lima M, Pybus O, Alcantara CJ (2015) Epidemiology of chikungunya virus in Bahia, Brazil, 2014–2015. PLoS Curr Outbreaks. https://doi.org/10.1371/currents.outbreaks.c97507e3e48efb946401755d468c28b2
15. Finger F, Genolet T, Mari L, Constantin G, Magny D, Magloire N (2016) Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proc Natl Acad Sci USA 113:6421–6426
16. Finkenstädt BF, Grenfell BT (2000) Time series modelling of childhood diseases: a dynamical systems approach. Appl Stat 49:187–205
17. Flasche S, Jit M, Rodríguez-Barraquer I, Coudeville L, Recker M, Koelle K et al (2016) The long-term safety, public health impact, and cost-effectiveness of routine vaccination with a recombinant, live-attenuated dengue vaccine (Dengvaxia): a model comparison study. PLoS Med 13:e1002181
18. Funk S, Salathé M, Jansen VAA, Funk S, Salathe M (2010) Modelling the influence of human behaviour on the spread of infectious diseases: a review. J R Soc Interface 7:1247–1256
19. González MC, Hidalgo CA, Barabási A-L (2008) Understanding individual human mobility patterns. Nature 453:779–782
20. Hadinegoro SR, Arredondo-García JL, Capeding MR, Deseda C, Chotpitayasunondh T, Dietze R et al (2015) Efficacy and long-term safety of a dengue vaccine in regions of endemic disease. N Engl J Med 373:1195–1206
21. Johansson MA, Vasconcelos PFC, Staples JE (2014) The whole iceberg: estimating the incidence of yellow fever virus infection from the number of severe cases. Trans R Soc Trop Med Hyg 108:482–487
22. Kraemer MUG, Faria NR, Reiner Jr RC, Golding N, Nikolay B, Stasse S et al (2017) Spread of yellow fever virus outbreak in Angola and the Democratic Republic of the Congo 2015–16: a modelling study. Lancet Infect Dis 17:330–338
23. Kraemer MUG, Hay SI, Pigott DM, Smith DL, Wint GRW, Golding N (2016) Progress and challenges in infectious disease cartography. Trends Parasitol 32:19–29
24. Kraemer MUG, Perkins TA, Cummings DAT, Zakar R, Hay SI, Smith DL et al (2015) Big city, small world: density, contact rates, and transmission of dengue across Pakistan. J R Soc Interface 12:20150468

25. Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Barker CM et al (2015) The global distribution of the arbovirus vectors *Aedes aegypti* and *Ae. albopictus*. eLife 4:e08347
26. Kraemer MUG, Sinka ME, Duda KA, Mylne A, Shearer FM, Brady OJ et al (2015) The global compendium of *Aedes aegypti* and *Ae. albopictus* occurrence. Sci Data 2:150035
27. Laniado D, Volkovich Y, Scellato S, Mascolo C, Kaltenbrunner A (2017) The impact of geographic distance on online social interactions. Inf Syst Front. https://doi.org/10.1007/s10796-017-9784-9
28. Lenormand M, Picornell M, Cantú-Ros OG, Tugores A, Louail T, Herranz R et al (2014) Cross-checking different sources of mobility information. PLoS ONE 9:e105184
29. Leta S, Jibat T, De Clercq EM, Amenu K, Kraemer MUG, Revie CW (2018) Global risk mapping for major diseases transmitted by *Aedes aegypti* and *Aedes albopictus*. Int J Infect Dis 67:25–35
30. Lourenco J, De Lima MM, Faria NR, Walker A, Kraemer MUG, Villabona-Arenas CJ et al (2017) Epidemiological and ecological determinants of Zika virus transmission in an urban setting. eLife 6:e29820
31. Mahmud AS, Metcalf CJE, Grenfell BT (2017) Comparative dynamics, seasonality in transmission, and predictability of childhood infections in Mexico. Epidemiol Infect 145:607–625
32. McNeill G, Bright J, Hale SA (2017) Estimating local commuting patterns from geolocated Twitter data. EPJ Data Sci 6:24
33. Messina JP, Brady OJ, Pigott DM, Golding N, Kraemer MUG, Scott TW et al (2015) The many projected futures of dengue. Nat Rev Microbiol 13:230–239
34. Metcalf CJE, Bjornstad ON, Ferrari MJ, Klepac P, Bharti N, Lopez-Gatell H et al (2011) The epidemiology of rubella in Mexico: seasonality, stochasticity and regional variation. Epidemiol Infect 139:1029–1038
35. Metcalf CJE, Bjornstad ON, Grenfell BT, Andreasen V (2009) Seasonality and comparative dynamics of six childhood infections in pre-vaccination Copenhagen. Proc R Soc Lond B, Biol Sci 276:4111–4118
36. Mills HL, Riley S (2014) The spatial resolution of epidemic peaks. PLoS Comput Biol 10:e1003561
37. Perkins TA, Garcia AJ, Paz-Soldan VA, Stoddard ST, Reiner RC, Vazquez-Prokopec G et al (2014) Theory and data for simulating fine-scale human movement in an urban environment. J R Soc Interface 11:20140642
38. Perkins TA, Metcalf CJE, Grenfell BT, Tatem AJ (2015) Estimating drivers of autochthonous transmission of chikungunya virus in its invasion of the Americas. PLoS Curr. https://doi.org/10.1371/currents.outbreaks.a4c7b6ac10e0420b1788c9767946d1fc
39. Perkins TA, Paz-Soldan VA, Stoddard ST, Morrison AC, Forshey BM, Long KC et al (2016) Calling in sick: impacts of fever on intra-urban human mobility. Proc R Soc Lond B, Biol Sci 283:20160390
40. Perkins TA, Scott TW, Le Menach A, Smith DL (2013) Heterogeneity, mixing, and the spatial scales of mosquito-borne pathogen transmission. PLoS Comput Biol 9:e1003327
41. Pya N, Wood SN (2015) Shape constrained additive models. Stat Comput 25:543–559
42. R Core Team (2016) R: a language and environment for computing. R Foundation for Statistical Computing, Vienna
43. Reich NG, Shrestha S, King AA, Rohani P, Lessler J, Kalayanarooj S et al (2013) Interactions between serotypes of dengue highlight epidemiological impact of cross-immunity. J R Soc Interface 10:20130414
44. Reiner RC, King AA, Emch M, Yunus M, Faruque ASG, Pascual M (2012) Highly localized sensitivity to climate forcing drives endemic cholera in a megacity. Proc Natl Acad Sci USA 109:2033–2036
45. Reiner RC, Perkins TA, Barker CM, Niu T, Chaves LF, Ellis AM et al (2013) A systematic review of mathematical models of mosquito-borne pathogen transmission: 1970–2010. J R Soc Interface 10:20120921
46. Reiner RC, Stoddard ST, Scott TW (2014) Socially structured human movement shapes dengue transmission despite the diffusive effect of mosquito dispersal. Epidemics 6:30–36
47. Salje H, Lessler J, Berry IM, Melendrez MC, 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
48. Salje H, Lessler J, Endy TP, Curriero FC, Gibbons RV, Nisalak A et al (2012) Revealing the microscale spatial signature of dengue transmission and immunity in an urban population. Proc Natl Acad Sci USA 109:9535–9538
49. Salje H, Lessler J, Kumar K, Azman AS, Rahman MW, Rahman M (2016) How social structures, space, and behaviors shape the spread of infectious diseases using chikungunya as a case study. Proc Natl Acad Sci USA 113:13420–13425
50. Simini F, González MC, Maritan A, Barabási A-L (2012) A universal model for mobility and migration patterns. Nature 484:96–100
51. Simmons CP, Farrar JJ, Chau NVV, Wills B (2012) Dengue. N Engl J Med 366:1423–1432
52. Siraj AS, Oidtman RJ, Huber JH, Kraemer MUG, Brady J, Johansson MA et al (2017) Temperature modulates dengue virus epidemic growth rates through its effects on reproduction numbers and generation intervals. PLoS Negl Trop Dis 11:e0005797
53. Stevenson JC, Pinchoff J, Muleba M, Lupiya J, Chilusu H, Mwelwa I et al (2016) Spatio-temporal heterogeneity of malaria vectors in northern Zambia: implications for vector control. Parasites Vectors 9:510
54. Stoddard ST, Forshey BM, Morrison AC, Paz-Soldan VA, Vazquez-Prokopec GM, Astete H et al (2013) House-to-house human movement drives dengue virus transmission. Proc Natl Acad Sci USA 110:994–999
55. Stoddard ST, Morrison AC, Vazquez-Prokopec GM, Paz Soldan V, Kochel TJ, Kitron U et al (2009) The role of human movement in the transmission of vector-borne pathogens. PLoS Negl Trop Dis 3:e481
56. Vanlerberghe V, Gómez-Dantés H, Vazquez-Prokopec GM, Alexander N, Manrique-Saide P, Coelho G et al (2017) Changing paradigms in *Aedes* control: considering the spatial heterogeneity of dengue transmission. Pan Am J Public Health 41:1–6
57. Vazquez-Prokopec GM, Bisanzio D, Stoddard ST, Paz-Soldan V, Morrison AC, Elder JP et al (2013) Using GPS technology to quantify human mobility, dynamic contacts and infectious disease dynamics in a resource-poor urban environment. PLoS ONE 8:e58802
58. Wesolowski A, Buckee CO, Engø-Monsen K, Metcalf CJE (2016) Connecting mobility to infectious diseases: the promise and limits of mobile phone data. J Infect Dis 214:S414–S420
59. World Health Organization (WHO) (2009) Dengue: guidelines for diagnosis, treatment, prevention, and control. World Health Organization, Geneva
60. Zipf GK (1946) The $\frac{P_1 P_2}{D}$ hypothesis: on the intercity movement of persons. Am Sociol Rev 11:677–686