

GIS-BASED EARLY WARNING SYSTEM FOR PREDICTING HIGH-RISK AREAS  
OF DENGUE VIRUS TRANSMISSION, RIBEIRÃO PRETO, BRAZIL

By  
Ryan Marc Carney

A Thesis Presented to  
The Faculty of the Yale School of Public Health  
Yale University

In Candidacy for the Degree of  
Master of Public Health

2010

## Abstract

Dengue virus (DENV) is currently the most rapidly spreading vector-borne disease, with an estimated 50-100 million infections per year and 2.5 billion people – 40% of the world’s population – at risk of infection. Since 1990, the city of Ribeirão Preto (pop. 563,000), Brazil, has experienced DENV epidemics of increasing severity, including the largest epidemic to date in 2010 ( $\approx$ 13,000 cases as of April 22). However, there are no vaccines or treatments for DENV, and the only method for reducing morbidity and mortality is through control of the principal mosquito vector, *Aedes aegypti*. To inform surveillance and control efforts in Ribeirão Preto, a geographic information system (GIS) was created along with an address locator for geocoding dengue cases. The primary objective of this study was to investigate the utility of modifying a West Nile virus early warning system, used successfully in California to predict and prevent human cases, that models viral amplification using a localized Knox test and Monte Carlo simulation approach based on parameters of vector and host biology. Results from this study, which represents the first spatially explicit model that uses human cases to predict future dengue risk, indicate that the modified DYCAST system provided early and accurate identification of high-risk areas in Ribeirão Preto, including detection of what appears to be the cryptic interepidemic focus of transmission that later developed into the severe 2006 epidemic. During the study period, DYCAST predicted up to 90.3% (4,234/4,690) of cases, at a maximum mean of 66.3 days prior to onset of illness. Maximum sensitivity and specificity was 83.8% and 78.8%, respectively, and relative risk of DENV infection was  $>10x$  higher in cells identified as high risk. Additionally, model efficacy was retained and even enhanced by including unconfirmed dengue cases in the analysis, which has important implications for increasing the utility and applicability of the model. These findings suggest that the DYCAST system could be utilized prospectively and in real-time to identify areas at high risk of DENV transmission, in order to target mosquito control, surveillance, and public education campaigns in a timely, efficient, and cost-effective manner. Furthermore, and in a departure from previous studies, this risk model was implemented using free, open-source, and cross-platform software that could provide an inexpensive and scalable GIS solution for the surveillance and control of DENV – and potentially other infectious diseases – by Ribeirão Preto and other public health agencies in the future. Protocols for generating the spatial datasets and installing the various software components are also provided.

## **Acknowledgments**

I would particularly like to thank my academic and thesis advisor, Durland Fish, PhD (Yale University), for his valuable direction and insight, as well as my preceptor Benedito Antônio Lopes da Fonseca, MD, PhD, MPH (Universidade de São Paulo) and colleague André Markon, MPH (University of Michigan) for their exceptional assistance during my summer internship. I would like to thank my thesis reader Maria Diuk-Wasser, PhD (Yale University) and professor Ted Holford, PhD (Yale University) for helpful comments and feedback, and acknowledge the author of the open source DYCAST scripts, Alan McConchie, MS (University of British Columbia). For their assistance in providing data and information used in this study, I would like to thank Marcelo Raspa, Claudio Souza de Paula, MD, and Maria Luiza, MD, from Secretaria Municipal de Saúde; Antônio Pazin Filho, MD from Universidade de São Paulo; and Wilson Yong from Companhia de Desenvolvimento Economico de Ribeirão Preto.

Funding for this research was provided through a US Centers for Disease Control and Prevention Training Grant Master of Public Health Fellowship.

## Table of Contents

### 1. Introduction

a.	Statement of general problem addressed by the thesis.....	8–10
b.	Objectives.....	10–11
c.	Relevant studies.....	11–12

### 2. Methods

a.	Data.....	13–15
b.	Risk model	
i.	Software.....	15–16
ii.	Procedure.....	17–19
iii.	Calibration.....	19–20
iv.	Evaluation.....	20–21

### 3. Results

a.	Calibration.....	22–26
b.	Evaluation.....	26–29

### 4. Discussion

a.	Summary of findings.....	30
b.	Significance of findings, assessment of applicability to current theory and practice.....	31–36
c.	Limitations of study and findings.....	36–38
d.	Relevant recommendations.....	38–41

### 6. References.....

42–47

### 7. Appendices

I.	Protocol – software installation, modification, FAQ.....	48–57
II.	Protocol – geocoding.....	58–71
III.	Protocol – grid creation.....	72–75
IV.	Protocol – DYCAST, ArcMap analysis.....	76–82

## List of Tables

**Table 1.** Number of confirmed dengue cases, Ribeirão Preto

**Table 2.** DYCAST model parameters

**Table 3.** Sensitivity analysis of geocoding parameters, composite address locator

**Table 4.** Model results, 2005 Jul 1–2006 Jun 30, Ribeirão Preto

**Table 5.** Confusion matrices comparing number of cells that contained dengue case(s) and number of cells identified as high risk by models, 2005 Jul 1–2006 Jun 30, Ribeirão Preto

## List of Figures

**Figure 1.** Current mapping methods used in Ribeirão Preto for dengue surveillance and control.

**Figure 2.** Geocoding process in ArcView 9.3.0.

**Figure 3.** Schematic of free and open source software used to implement the DYCAST model.

**Figure 4.** Schematic of the dengue virus DYCAST procedure.

**Figure 5.** Average monthly temperatures, Ribeirão Preto, 2003-2009.

**Figure 6.** Dengue cases in Ribeirão Preto per week, by year.

**Figure 7.** Map illustrating dengue cases (black circles) and high-risk cells generated at analysis thresholds 1–15 for Jan 31, 2006, Ribeirão Preto.

**Figure 8.** Graph illustrating number of high-risk cells, maximum p-value, and proportion of p-values  $\leq \alpha$  (0.1) for high-risk cells generated at analysis thresholds 1–15 for Jan 31, 2006, Ribeirão Preto.

**Figure 9.** Receiver operating characteristic (ROC) plot for models D10, D05, and D10'.

**Figure 10.** Number of confirmed dengue cases and DYCAST cells identified as high risk by models D10 (red), D05 (orange), and D10' (yellow) per day, Ribeirão Preto, Jul 1, 2005–Jun 30, 2006.

**Figure 11.** Confirmed dengue cases and DYCAST risk maps displaying high-risk cells as identified by models D10 (red), D05 (orange), and D10' (yellow), Ribeirão Preto.

**Figure 12.** Number of suspected cases confirmed (red) and unconfirmed (green) for dengue infection per year, 1998-2009, Ribeirão Preto.

## Body of the Thesis

i. **Thesis Type:** *Research Study Demonstrating Mastery of Methodology*

ii. **Article Summary Line**

The DYCAST early warning system was effective at identifying areas at high risk of dengue virus transmission during epidemic and interepidemic periods.

iii. **Running Title**

Dynamic early warning system for dengue virus

iv. **Keywords**

arboviruses, Brazil, dengue virus, disease outbreaks, geographic information systems, humans, insect vectors, mosquitoes, surveillance, transmission

## Introduction

### Statement of general problem addressed by the thesis

Dengue virus (DENV; genus *Flavivirus*, family *Flaviviridae*) is currently the fastest-spreading vector borne disease, with an estimated 50-100 million infections per year and 2.5 billion people – 40% of the world’s population – at risk of infection (1,2). Exacerbated by growing trends of urbanization and environmental perturbation, this tremendous global disease burden falls almost exclusively on developing countries, which generally lack the resources and infrastructure for sufficient surveillance and control of the virus. While most DENV illnesses are asymptomatic or develop into the mild form of the disease, classic dengue fever (DF), approximately 3% of illnesses develop into dengue hemorrhagic fever (DHF) and dengue shock syndrome (DSS), with mortality rates over 10% (3-5). These severe forms of illness are due to antibody-dependent enhancement (ADE) caused by subsequent infection with one of the other four dengue serotypes (DENV-1, -2, -3, and -4). However, there are no vaccines or treatments available for dengue, therefore the only method for reducing human morbidity and mortality is through control of the principal mosquito vector, *Aedes aegypti*. Thus, there is an acute need for an early warning system to target such prevention and control efforts, and in particular, to identify areas where the virus is cryptically maintained during interepidemic periods (6).

From 1990-1991, the first dengue epidemic occurred in Riberão Preto (652 km<sup>2</sup>; population: 563,000), a northern city in the State of São Paulo, Brazil. During this outbreak there were 8,900 reported cases (7), and a survey conducted in 1992 detected



5.4% seroprevalence for DENV-1 IgG antibodies among participating residents (8). DENV-2 and -3 were detected in 1998 and 2001, respectively, with the latter triggering a large epidemic during that same year (Benedito Antônio Lopes da Fonseca, University of São Paulo, pers. comm.). Currently, DENV-3 is the predominant serotype within Riberão Preto, which has experienced epidemics of increasing severity (Table 1), including the largest epidemic to date in 2010 ( $\approx$ 13,000 confirmed cases as of April 22; 9). Vector control efforts are initiated in the city once the monthly incidence rate exceeds the average rate from the previous 10 non-epidemic years (Claudio Souza de Paula, Secretaria Municipal de Saúde, pers. comm.). However, this practice only provides a temporal threshold for epidemic control, and ignores the spatial aspects of dengue transmission. Complicating surveillance and control efforts is the fact that the health departments and vector control agency have relied exclusively on paper-based mapping of cases and vector data (Figure 1a,b).

Table 1. Number of confirmed dengue cases, Riberão Preto

Year	Total cases	Classic dengue	Dengue with complications	DHF	DSS
<b>2010*</b>	<b>12,933</b>	-	-	<b>31</b>	-
<b>2009</b>	<b>1,685</b>	<b>1,671</b>	<b>5</b>	<b>9</b>	<b>0</b>
2008	1,073	1,066	3	2	2
<b>2007</b>	<b>2,752</b>	<b>2,733</b>	<b>15</b>	<b>4</b>	<b>0</b>
<b>2006</b>	<b>6,051</b>	<b>6,026</b>	<b>11</b>	<b>13</b>	<b>1</b>
2005	645	639	1	5	0
2004	50	50	0	0	0
2003	808	800	0	8	0
2002	360	360	0	0	0
<b>2001</b>	<b>3,217</b>	<b>3,216</b>	<b>0</b>	<b>1</b>	<b>0</b>
2000	211	211	0	0	0
1999	320	320	0	0	0
1998	268	268	0	0	0
Total	17,440	17,360	35	42	3

**Boldface** denotes years of critical transmission, defined by the Brazilian Ministry of Health as incidence rates  $>300$  per 100,000 (de Paula, pers. comm.)

\*As of April 22, 2010 (9)



Figure 1. Current mapping methods used in Ribeirão Preto for dengue surveillance and control. a. color-coded push pins used to plot dengue cases in one of the health districts (2007), and b. one of dozens of papers comprising a map of the city used to track and plan for mosquito surveillance and control.

## Objectives

The objectives of this study were threefold. First, this study aimed to create a comprehensive GIS that could be utilized by local agencies in Ribeirão Preto for the surveillance and control of dengue, and in particular, a means to geocode the addresses of dengue cases for mapping purposes. Second, these spatial datasets would be used to implement a dengue risk model, which would be evaluated for use as a prospective, early warning system in Ribeirão Preto. The third objective was to use this model to detect cryptic transmission foci responsible for interepidemic maintenance of the virus during

the dry season, which would be valuable for targeting mosquito control efforts to interrupt transmission prior to reemergence of epidemic amplification.

### **Relevant Studies**

The majority of research on DENV transmission has focused on the epidemic periods of the disease, ignoring interepidemic transmission that nonetheless contributes considerable disease burden through productivity loss and absenteeism (1). What is known is that this phenomenon appears to be driven more by intrinsic population dynamics of the vector than extrinsic factors such as climate (6), suggesting that modeling efforts should concentrate more on intra-seasonal instead of inter-seasonal factors. However, in addition to numerous methodological shortcomings of using immature mosquito indices to predict risk of human infection (10), the vector surveillance data collected in Ribeirão Preto is prohibitive for using existing dengue transmission models (11,12) due to the entomological stage sampled (larval instead of pupal) and spatial scale at which data is aggregated (data not shown). Other researchers have used various spatial and spatiotemporal methods to characterize distributions of human cases (13-16), however all of these models have been only descriptive in nature and do not provide any statistical means for assessing future risk of dengue transmission.

However, previous studies have shown the Dynamic Continuous-Area Space-Time (DYCAST) model to be successful as an early warning system for the related mosquito-borne flavivirus, West Nile virus (WNV). DYCAST identifies area at high risk of WNV transmission to humans by using statistical analysis of public reports of dead birds, based on biological parameters of *Culex* spp. mosquito vectors and avian hosts.

This model was first implemented, retrospectively, in New York City in 2000 and Chicago in 2001, and implemented prospectively in California since 2005 as an early warning system (17-19). In California, DYCAST has been used to assist mosquito larviciding and adulticiding efforts (20), including emergency aerial mosquito control conducted during an unprecedented epidemic in 2005, which has been shown to have reduced human morbidity and potentially mortality from WNV infection (21). Given these successful results implementing DYCAST as an early warning system for a related arboviral disease, this model was selected for the present study and adapted for use as a dengue risk model.

## Methods

### Data

Excel (Microsoft Corporation, Redmond, WA, USA) databases containing records of all suspected dengue cases (n=53,256) in Ribeirão Preto from January 1, 1998 through January 25, 2010 were obtained from the Ribeirão Preto Municipal Health District's Division of Epidemiological Surveillance (Divisão de Vigilância Epidemiológica. Secretaria Municipal de Saúde de Ribeirão Preto). Information included residential address, date of onset of illness, and clinical diagnosis of confirmed (n=17,440; Table 1) and unconfirmed (n=35,816) dengue cases. These data were scrubbed and integrated into a single Excel database, and the confirmed cases used to plot case incidence by week in order to assess seasonality of epidemic and interepidemic periods. Locations of all suspected cases were mapped by using ArcView 9.3.0 (Environmental Systems Research Institute, Inc.; ESRI, Redlands, CA, USA) and a composite address locator to geocode records based on street address, Código de Endereçamento Postal (CEP; zip code), and bairro (district) (Figure 2a); this locator was built by using a line shapefile containing street reference data for Ribeirão Preto and a Brazilian-formatted locator index, both obtained from ESRI (see Appendix). One source of spatial inaccuracy in geocoding address data is the offset value used to specify distance of residences from the centerline of their respective street, which can vary greatly between rural and urban areas. Thus, an appropriate offset was derived using empirical data, which consisted of 11,909 suspected dengue cases in Ribeirão Preto, from approximately January 1, 2007 through May 31, 2008, that had been previously

georeferenced to centroids (center) of residential lots by Companhia de Desenvolvimento Economico de Ribeirão Preto (CODERP; Wilson Yong, unpub. data). The median distance between cases and their respective street centerline was then obtained (Figure 2b; see Appendix); cases >100 m from a street centerline were excluded, as these represented addresses not included in the street reference data. The mean offset value was used in all subsequent geocoding processes.

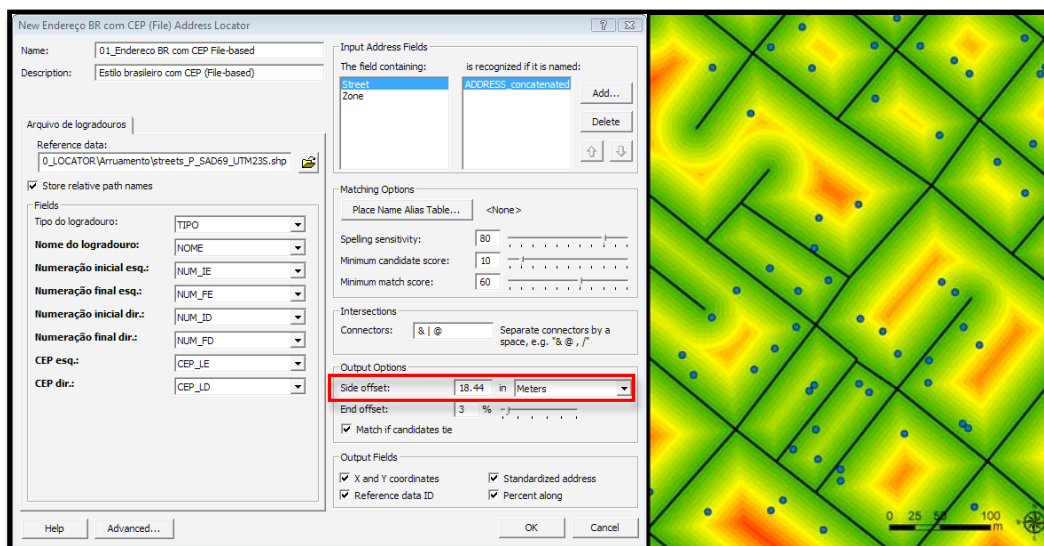


Figure 2. Geocoding process in ArcView 9.3.0. a. Código de Endereçamento Postal (CEP; zip code)-based address locator, illustrating matching parameters (“Matching Options”) and side offset (red box); b. raster layer of distances from street centerline (from green to red) sampled by cases (blue circles) to calculate median offset.

Selection of address locator matching parameters (spelling sensitivity, minimum candidate score, and minimum matching score) was made through a step-wise sensitivity analysis. This included comparing geocoding results from all suspected cases, using both

individual address locators and 14 combinations of parameters with the composite locator (see Appendix). An optimal parameter set was selected by examining results for accuracy. Approximately half (n=7,500) of the unmatched records were then manually rematched; findings from this process were then used to automatically rematch the remainder. Pearson correlation was used to test whether the percentage of successfully matched addresses changed over time, which might indicate changes in data quality or the presence of addresses in areas not represented by street reference data.

## **Risk model**

### **Software**

In previous studies, the DYCAST model was implemented by using expensive and specialized GIS software, Smallworld 3.2.1 (General Electric Company, Fairfield, CT, USA), and its proprietary Magik computer programming language (17-19). More recently, the model has been recreated by using the widely used computer programming language, Python ([www.python.org](http://www.python.org)), which provides for the implementation of the model using free, open source, and cross-platform (Windows and Mac operating systems) software (Alan McConchie, University of British Columbia, unpub. data). The present study implemented this open source version of the model on a 64-bit Windows Vista Business Edition platform (Microsoft Corp.). DYCAST scripts were modified in order to run the model using dengue-specific parameters and the spatial reference appropriate for the study region (see Appendix); scripts were executed using Python 2.6. Psychopg 2.0.14 ([www.initd.org/psycopg](http://www.initd.org/psycopg)) provided porting to PostgreSQL 8.4.3 object-relational database management system ([www.postgresql.org](http://www.postgresql.org)); a graphical user interface (GUI) was

provided by pgAdminIII (Figure 3). Support for geographic objects was added by using PostGIS 1.5.1 ([www.postgis.org](http://www.postgis.org)) and shp2pgsql graphical loader plugin. The model's analysis grid, risk maps, and animations were created by using ArcView 9.3.0, along with its Spatial Analyst and Tracking Analyst extensions (see Appendix). While ArcView is commercial software, these operations could be accomplished in future implementations by using the free and open source GIS software, Geographic Resources Analysis Support System GIS (GRASS GIS; [grass.osgeo.org](http://grass.osgeo.org)).

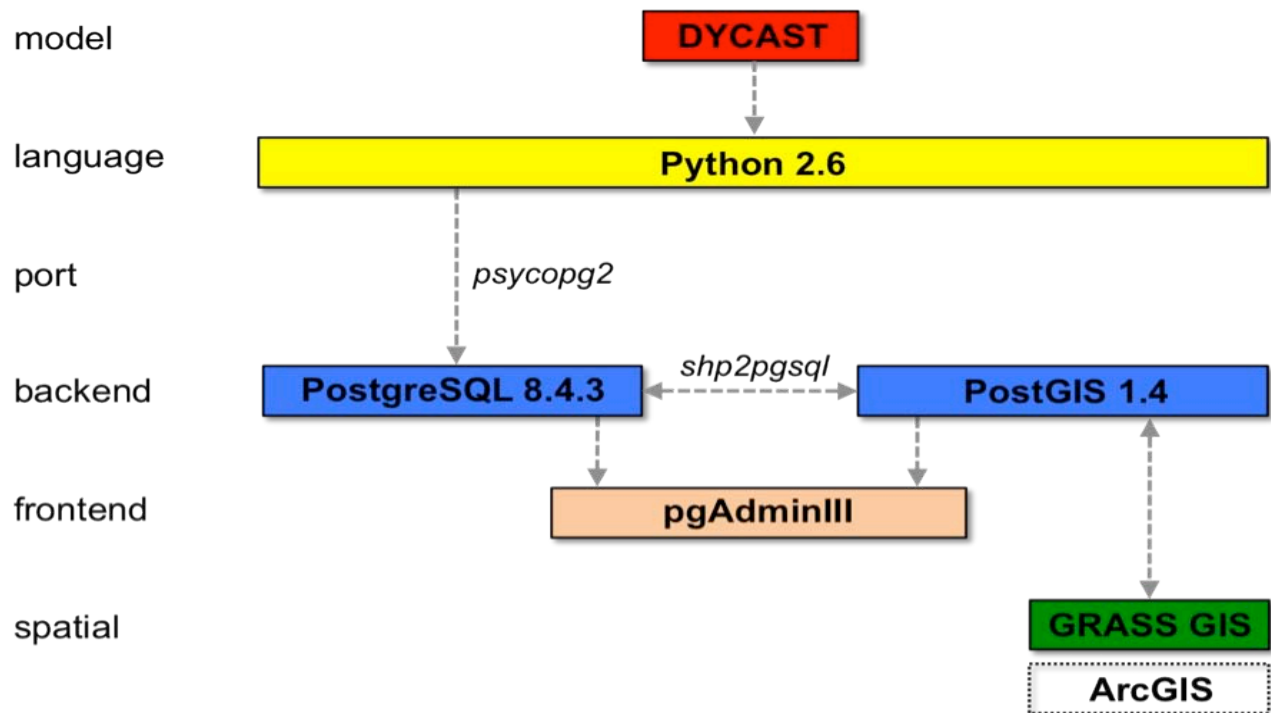


Figure 3. Schematic of free and open source software used to implement the DYCAST model. Commercial ArcGIS software (i.e., ArcView) was used in this study to create spatial datasets; however, this could be accomplished in future implementations by using free and open source GIS software, GRASS GIS.



## Procedure

A digital map of Ribeirão Preto was superimposed with a grid consisting of 71,824 1-hectare (0.01 km<sup>2</sup>) cells. A localized Knox test (22-23) was implemented from the centroid of each cell for which the number of dengue cases within the spatial and temporal domains met or exceeded the analysis threshold (see below). The radius of the spatial domain was based on a 600 m maximum flight range of *Ae. aegypti* (24,25), a distance which fell within estimates of 200 m (26) and 800 m (27) derived from suburban and urban field studies in Rio de Janeiro, Brazil using laboratory-raised mosquitoes. The 28-day temporal domain represented one complete human-vector-human infection cycle (Figure 4); this accounted for a 10-day extrinsic incubation period of DENV in *Ae. aegypti* (28-30) and two human infection cycles of 9 days each, which consist of a 6-day intrinsic incubation period (31-32) and a 3-day infectious symptomatic period (32). These bounds define the spatiotemporal domain, within which cases were analyzed for close pairing in space and time. Closeness in space was defined as 100 m, to encompass the mean flight range of *Ae. aegypti* (26,24,33,34). Closeness in time was defined as 4 days, and represented the length of the *Ae. aegypti* gonotrophic cycle based on field (35,36) and laboratory (33,37) findings at temperatures consistent with those in Ribeirão Preto (Figure 5). Statistical significance of case pairing was assessed by using unconditional Monte Carlo simulations (18,38). P-values denote the probability that clustering in space and time is caused by random chance; high-risk was defined at  $\alpha=0.1$  in accordance with previous DYCAST models (17-19). This procedure was repeated at each centroid to create an interpolated surface of risk. Parameters used in the WNV and DENV models are shown in Table 2 for comparison purposes.

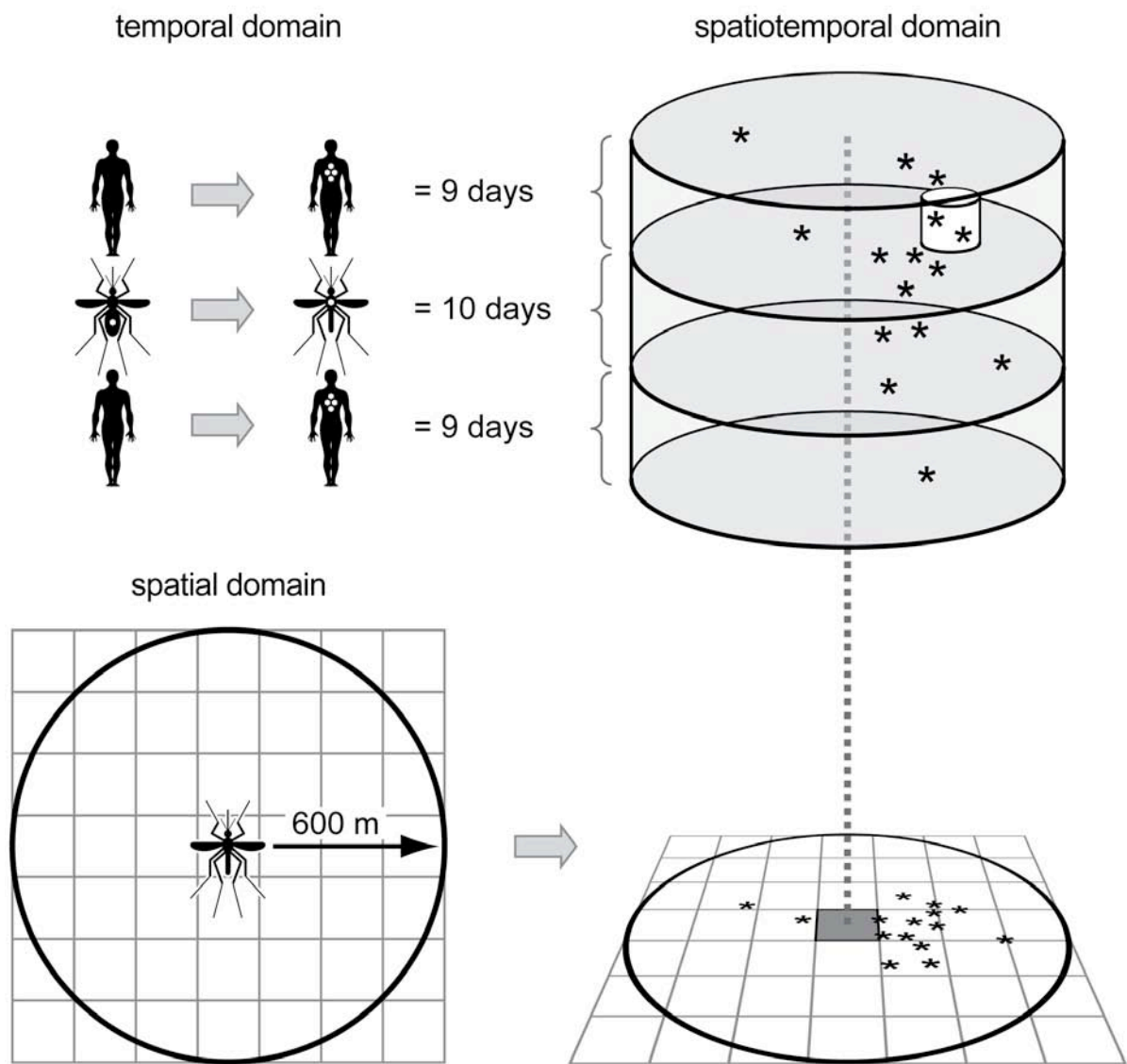


Figure 4. Schematic of the dengue virus DYCAST procedure, illustrating domains of localized Knox test implemented at the centroid of an individual 1-hectare grid cell (grid not to scale). The radius of the  $\approx 113$ -hectare ( $1.13 \text{ km}^2$ ) spatial domain is based on the maximum flight range of *Aedes aegypti*, and the 28-day temporal domain accounts for the extrinsic incubation period of DENV and two human infection cycles. These bounds define the spatiotemporal domain, within which cases (asterisks) were analyzed for close pairing in space (100 m) and time (4 days) (small white cylinder). Statistical significance of case pairing was assessed by using unconditional Monte Carlo simulations ( $\alpha=0.1$ ).

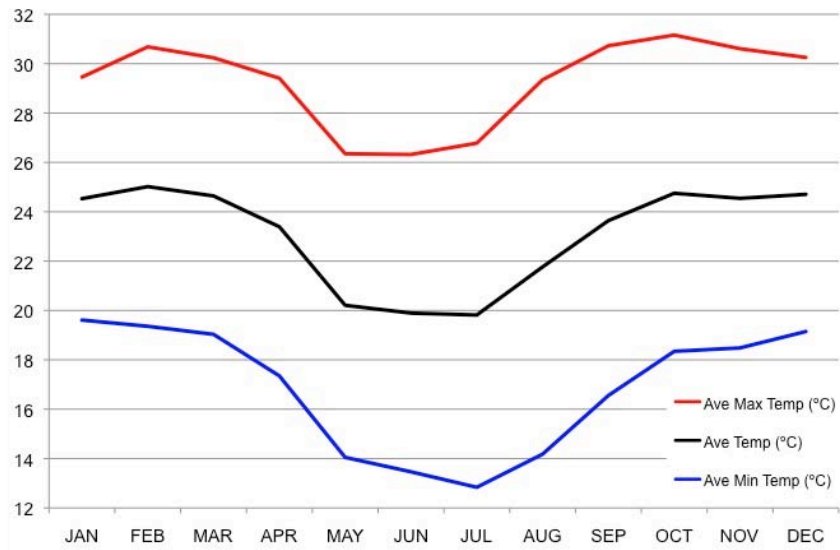


Figure 5. Average monthly temperatures, Ribeirão Preto, 2003-2009. Data source: Centro Integrado e Informações Agrometeorológicas (<http://www.ciagro.sp.gov.br/ciagroonline/Listagens/MonClim/LMClmLocal.asp>)

Table 2. DYCAST model parameters\*

Virus model	Spatial domain (m)	Spatial closeness (m)	Temporal domain (days)	Temporal closeness (days)	Threshold (number reports)
West Nile	2,410	402	21	3	15
dengue	600	100	28	4	10,5

\*West Nile virus parameters taken from (19).

## Calibration

In order to determine an appropriate analysis threshold value, a sensitivity analysis was conducted for a date representative of an incipient epidemic (January 31, 2006). This included examinations of high-risk areas and p-value distributions generated by the DYCAST model at thresholds of 1 through 15. Optimal threshold(s) were subsequently used to run risk for the duration of the study period, for which year 2006 was selected as it represented the largest dengue epidemic in Ribeirão Preto to date

(excluding the current 2010 epidemic). The preceding interepidemic period in 2005 was also included in the study period, the exact date of which was determined by examination of case incidence to validate current temporal definitions. In Ribeirão Preto, the interepidemic period is officially defined as July 1 through October 31 of each year, approximately weeks 27 through 43 (Maria Luiza, Secretaria Municipal de Saúde de Ribeirão Preto, pers. comm.); ecologically, this coincides with the rainy season that begins around November (de Paula, pers. comm).

## **Evaluation**

Results from the risk models were analyzed by using ArcMap, Excel:mac 2008 version 12.2.4, and StatPlus:mac LE 2009 version 5.8.0.0 (AnalystSoft, Inc., Vancouver, BC, Canada). Prediction was defined as the identification of a cell as high risk prior to or on the date of onset of illness of the earliest confirmed case located within that cell. If a cell was identified as high risk after the date of onset of illness, or the cell was never identified as high risk and a case occurred within it, it was designated false negative. Out of the total area analyzed, cells considered to be potentially at risk were those within 618.44 m of street reference data. This distance was based on the model's 600 m spatial domain plus the 18.44 m geocoding side offset (see below), and represented the maximum distance from the street centerline that a cell could be identified as high-risk. This provided for a more conservative estimate of the true negative rate (and thus validity indices) by eliminating areas of the city where cases could not be mapped, and resulted in a study region of 25,487 cells, a total of 254.9 km<sup>2</sup>.

Results were used to calculate metrics of model efficacy, such as prediction rates and number of days prior to onset of illness that cases were predicted. Confusion matrices were also constructed to compare number of cells that contained dengue cases and number of cells identified as high risk by each model. These were used to generate measures of validity and association, including sensitivity, specificity, Youden's J, positive (PPV) and negative predictive values (NPV), classification accuracy, relative risk (RR), and kappa statistic of chance-adjusted agreement (39). Because kappa is highly sensitive to both prevalence and classification bias, a prevalence-adjusted, bias-adjusted kappa statistic was used (PABAK; 40). Results were also used to generate receiver operating characteristic (ROC) plots, and the area under the ROC curve (AUC) was estimated nonparametrically by using the Wilcoxon rank-sum statistic (41).

## Results

### Calibration

Results of weekly incidence plotted by year confirm the validity of the July 1 through October 31 temporal bounds of the interepidemic period; specifically, week 44 corresponds to increased incidence rates in 2005 and 2009 (Figure 6, inset), which preceded the unprecedented epidemics of 2006 and 2010. Thus, July 1, 2005 through June 30, 2006 was selected as the duration for the model study period.

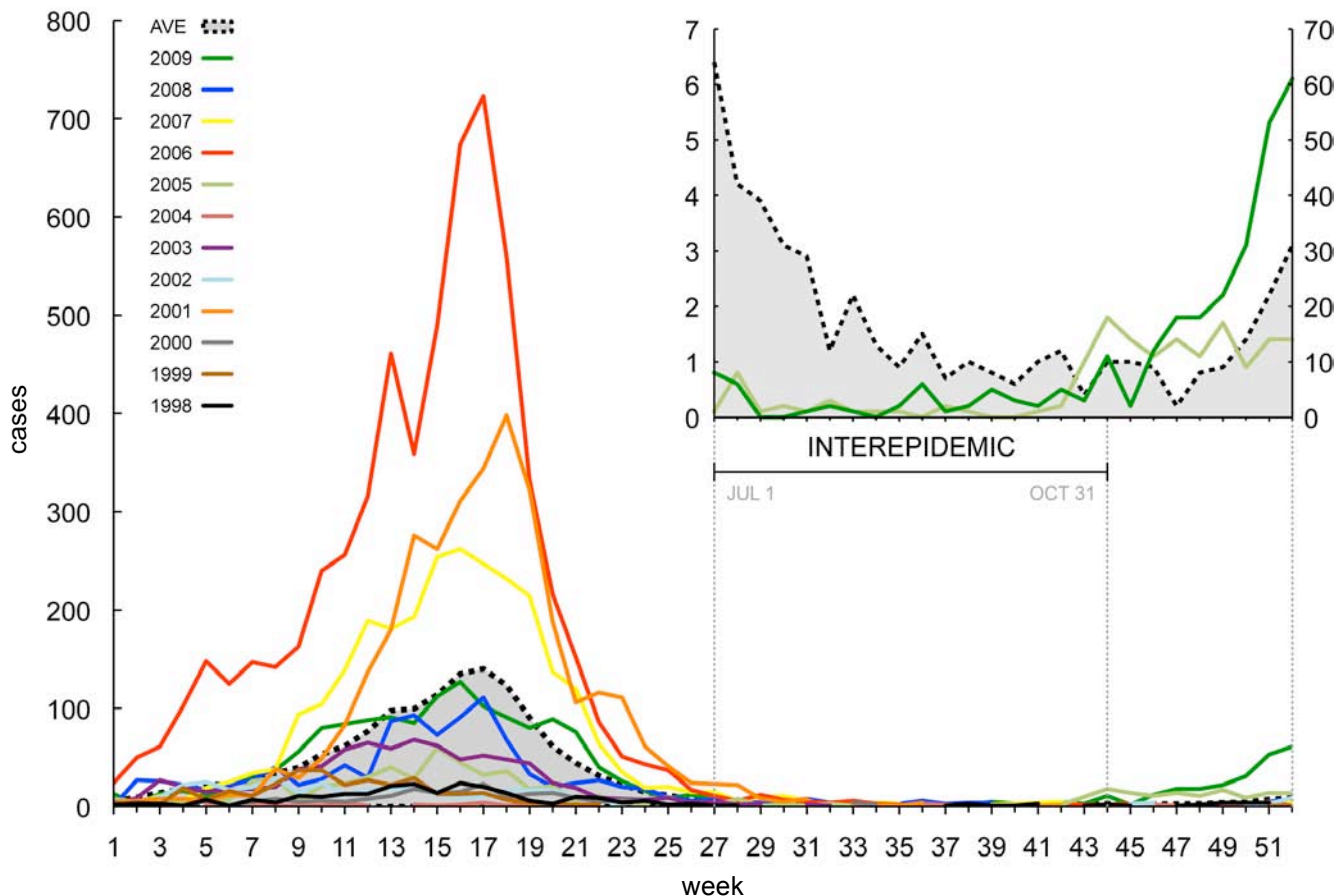


Figure 6. Dengue cases in Ribeirão Preto per week, by year. Average is the mean case incidence over the 12 years shown; average in the inset does not include 2005 and 2009 data, however. Inset x-axis has the same scale, Jul 1–Dec 31; inset y-axes have the same units (cases).

Geocoding offset analysis yielded a mean distance of 18.44 m (60.50 feet) from residential lot centroid to street centerline; this value was used as the side offset value in the subsequent geocoding of dengue cases (Figure 2a). The composite locator provided a 1.0% increase in matched records compared to geocoding using the CEP locator alone. Findings from the sensitivity analysis indicated that out of the 14 different combinations of matching parameters, 60, 10, and 23 was the optimal combination of spelling sensitivity, minimum candidate score, and minimum match score, respectively (Table 3, see Appendix). Lowering the spelling sensitivity below 60 increased the number of inaccurate matches, due to a greater number of tied records (which are considered matches) that would otherwise be classified as unmatched. Subsequent manual examination and rematching of unmatched records indicated that the optimal parameter combination should be modified further to 60, 10, and 10, respectively. This rematching process also increased the total number of matched records by 1.2%, to 73.2%. Results

Table 3. Sensitivity analysis of geocoding parameters, composite address locator

Parameters *	% total matched	Number matched	Number tied	Number unmatched
80-10-60†	57.4%	30,006 (56%)	589 (1%)	22,661 (43%)
80-10-60‡	14%	7,488 (14%)	124 (0%)	45,644 (86%)
80-10-60	58%	30,503 (57%)	598 (1%)	22,155 (42%)
40-10-60	59%	30,958 (58%)	606 (1%)	21,692 (41%)
40-10-30	68%	35,866 (67%)	777 (1%)	16,613 (31%)
80-10-23	64%	35,732 (62%)	903 (2%)	16,621 (31%)
<b>60-10-23</b>	<b>72%</b>	<b>37,544 (70%)</b>	<b>940 (2%)</b>	<b>14,772 (28%)</b>
50-10-23	74%	38,485 (72%)	1,071 (2%)	13,700 (26%)
40-10-23	76%	39,114 (73%)	1,558 (3%)	12,584 (24%)
20-10-23	81%	39,474 (74%)	3,529 (7%)	10,253 (19%)
40-10-15	77%	39,420 (74%)	1,573 (3%)	12,263 (23%)
20-10-15	82%	39,800 (75%)	3,551 (7%)	9,905 (19%)
20-5-15	82%	39,800 (75%)	3,551 (7%)	9,905 (19%)
20-5-5	82%	40,073 (75%)	3,568 (7%)	9,615 (18%)
5-5-5	84%	39,893 (75%)	4,618 (9%)	8,745 (16%)
0-0-0	88%	40,352 (76%)	6,490 (12%)	6,414 (12%)

**boldface** denotes optimal parameters used in final dataset

\*Spelling sensitivity, Minimum candidate score, and Minimum match score, respectively.

Default values are 80-10-60.

†CEP (zip code)-based address locator (individual)

‡Bairro (district)-based address locator (individual)

from the Pearson correlation test indicated no significant linear correlation between percentage of geocode matches and year ( $p=0.1174$ ,  $R=-0.47644$ ).

Results from the threshold sensitivity analysis indicated the occurrence of false positives and pair dependencies at low threshold numbers, which is an inherent statistical artifact of the Monte Carlo distributions. At a threshold of 2, for example, high-risk cells were predicted around every pair of cases (Figure 7). However, this effect was eliminated at thresholds  $\geq 10$ , which resulted in all high-risk cell  $p$ -values  $\leq \alpha$  of 0.1 (Figure 8). Thus, a threshold of 10 was used for the initial risk model, D10. Additionally, in order to investigate the utility of using a lower threshold to increase sensitivity of the model during interepidemic periods, a threshold of 5 was used for a second risk model, D05. This threshold was based on an acceptable proportion of high-risk cell  $p$ -values  $\leq \alpha$  (88.1%; 998 of 1,133 cells), as well as qualitative examination of the risk maps, which indicated a sufficient balance of over- and under-prediction of risk at this threshold (Figure 7, Figure 8). A third model (D10<sup>?</sup>) was also run which used as its input all suspected cases of dengue, which included 4,496 unconfirmed cases in addition to the 4,960 confirmed cases used in the other models. The purpose of this iteration was to evaluate the effect that using raw syndromic surveillance data would have on the model's efficacy.



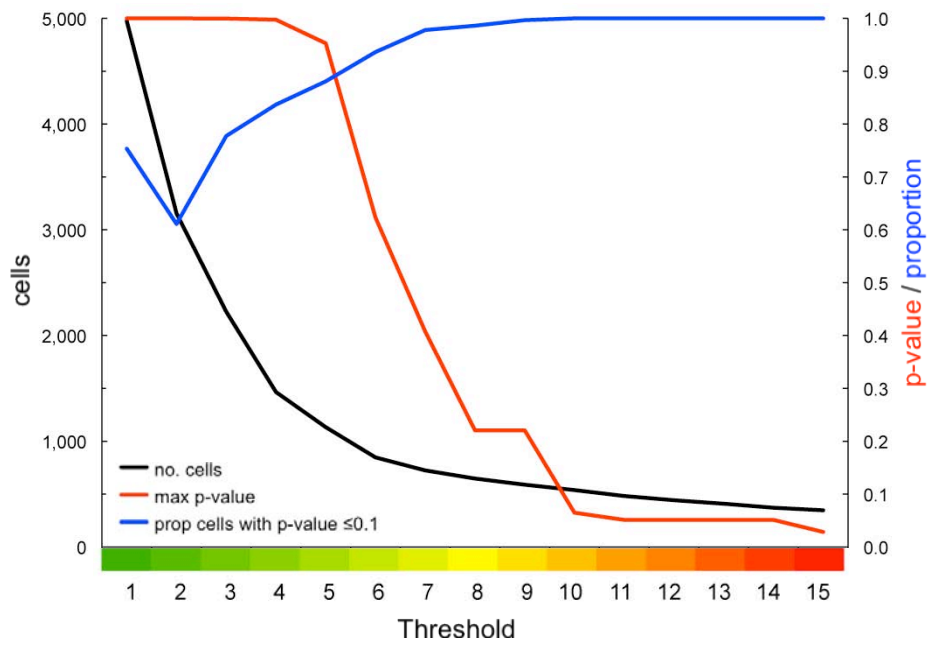
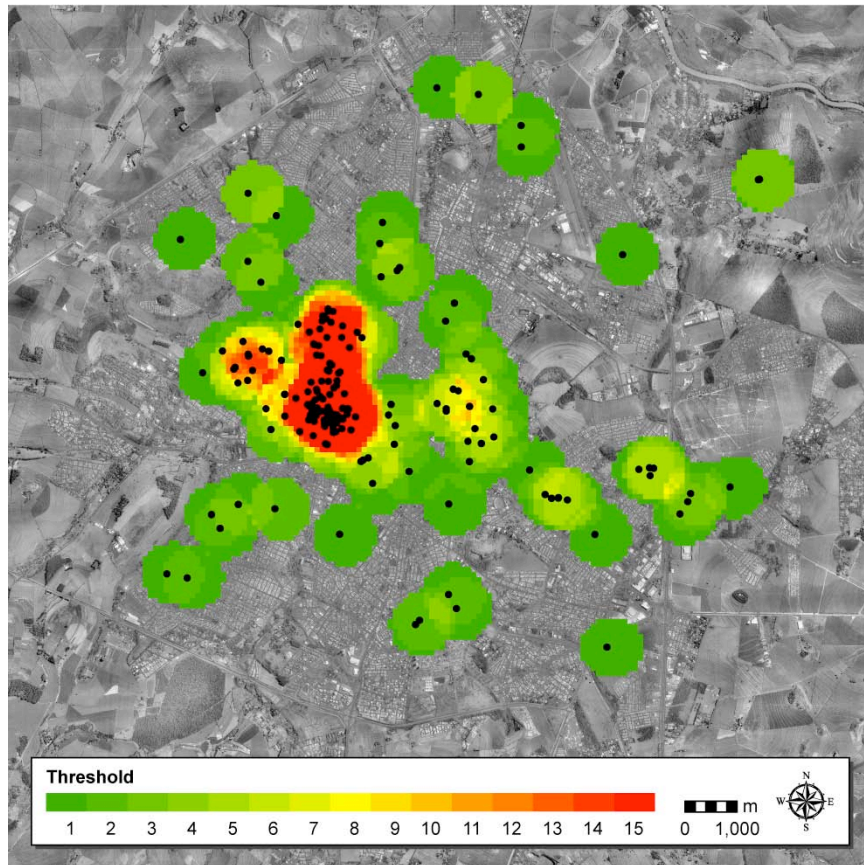


Figure 7. Map illustrating dengue cases (black circles) and high-risk cells generated at analysis thresholds 1–15 for Jan 31, 2006, Ribeirão Preto. Cases displayed are those with an onset of illness between Jan 3–Jan 31; this corresponds to the 28-day temporal domain of the model and thus includes all cases contributing to risk at that time.

Figure 8. Graph illustrating number of high-risk cells, maximum p-value, and proportion of p-values  $\leq \alpha$  (0.1) for high-risk cells generated at analysis thresholds 1–15 for Jan 31, 2006, Ribeirão Preto.

## Evaluation

Models D10', D05, and D10 first identified cells as high-risk on July 6, 14, and October 31; overall, cells were identified as high risk for a mean total of 92.3, 83.8, and 65.4 days, respectively (Table 4). During the study period there were 4,690 confirmed dengue cases, of which a maximum 90.3% (4,237/4,690) was predicted by the models at a mean of 66.3 days prior to onset of illness (D10'). Validity indices derived from confusion matrices (Table 5) yielded maximum sensitivity of 83.8% (D10') and maximum specificity of 78.8% (D10); model D10' exhibited the highest Youden's J statistic, 0.561. Both measures of agreement, accuracy (0.777) and kappa (0.554), as well as PPV (24.2%), were highest in D10; NPV was highest in D10' (97.8%). All three RR values differed significantly from unity, with a maximum of 10.569 (95% CI 9.498–11.760,  $p < 0.0001$ ) for D10'. ROC analysis yielded an AUC of 0.783 (Figure 9); given the similar results obtained by D05 and D10', exclusion of the latter from the calculation yielded a virtually equivalent AUC of 0.780.

Table 4. Model results, 2005 Jul 1–2006 Jun 30, Ribeirão Preto

Cells	Model*		
	D10	D05	D10'
Date of first high-risk cell	Oct 31	Jul 14	Jul 6
% high-risk cells (out of 25,487)	27.7%	34.9%	34.0%
No. high-risk cells	7,062	8,899	8,660
No. days cells identified high risk			
mean	65.4	83.8	92.3
std dev	38.9	46.1	47.8
median	62	83	95
max	199	242	243
<b>Cases</b>			
Prediction rate (out of 4,690)	79.1%	90.3%	<b>90.3%</b>
No. cases predicted	3,710	4,234	<b>4,237</b>
No. days predicted prior to onset			
mean	48.0	62.5	<b>66.3</b>
std dev	35.6	41.4	41.0
median	42	57	61
max	234	304	315
<b>Validity, association†</b>			
Sensitivity	67.1%	83.1%	<b>83.8%</b>
Specificity	<b>78.8%</b>	71.3%	72.3%
Youden's J	0.458	0.544	<b>0.561</b>
Accuracy	<b>0.777</b>	0.724	0.734
Kappa (PABAK)‡	<b>0.554</b>	0.447	0.467
Positive predictive value	<b>24.2%</b>	22.6%	23.4%
Negative predictive value	95.9%	97.7%	<b>97.8%</b>
Relative risk	5.969	9.704	<b>10.569</b>
95% CI, lower limit	5.502	8.735	9.498
95% CI, upper limit	6.476	10.779	11.760

**Boldface** denotes optimal value of the three models.

\*TH10 and TH05 used confirmed dengue cases (n=4,690) and a threshold of 5 and 10, respectively; TH10' used suspected dengue cases (n=9,186; both confirmed and non-confirmed cases) and a threshold of 10.

†Calculated from confusion matrix (Table 5; 39).

‡Prevalence-adjusted, bias-adjusted kappa (40).

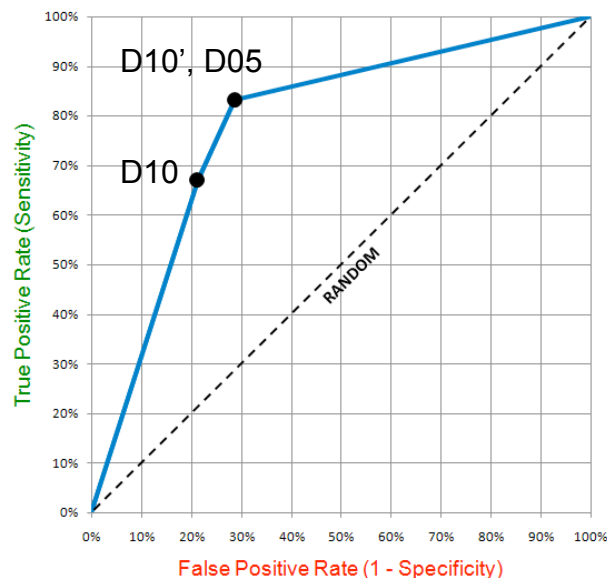


Figure 9. Receiver operating characteristic (ROC) plot for models D10, D05, and D10'. Area under the ROC curve (AUC) is equal to 0.783.

Table 5. Confusion matrices comparing number of cells that contained dengue case(s) and number of cells identified as high risk by models, 2005 Jul 1–2006 Jun 30, Ribeirão Preto\*

	Model*											
	D10			D05			D10'†					
	Contained case		Total	Contained case		Total	Contained case		Total			
High risk	Yes	1,568		4,915	6,483		Yes	1,944		6,647	8,591	Yes
	No	770	18,234	19,004	No	394	16,502	16,896	No	379	16,737	17,116
		2,338	23,149	25,487		2,338	23,149	25,487		2,338	23,149	25,487

\* True positive (Yes/Yes) designates cell identified by DYCAST as high risk prior to or on the date of onset of illness of earliest case located within cell. If cell was identified as high risk after date of onset of illness, or cell was never identified as high risk and a case occurred within it, it was designated false negative (Yes/No). Number of cells that contained cases is less than the number of confirmed cases (4,690) due to 50.1% of cases occurring in cells containing other cases.

†TH10 and TH05 used confirmed dengue cases (n=4,690) and a threshold of 5 and 10, respectively; TH10' used suspected dengue cases (n=9,186; both confirmed and non-confirmed cases) and a threshold of 10

‡Denotes results from model using a threshold of 10 and all suspected dengue cases (n=9,186) instead of only confirmed cases.

One interesting finding is that D10' detected a cluster of cases in the western portion of the city at the beginning of the interepidemic period, identifying high-risk cells from July 6–August 8, 2005 (D05 also detected this cluster, but only identified risk between July 14–17) (Figure 10). What is important about this result is that this cluster was in the same location as the sole cluster of incipient risk identified in late October of that same year (Figure 11). Examination of the daily risk maps suggests that this October cluster was the initial focus of transmission for the 2006 epidemic, appearing to spill over into the surrounding areas (see Supplemental Files for an animation of D10' risk maps during entire study period).

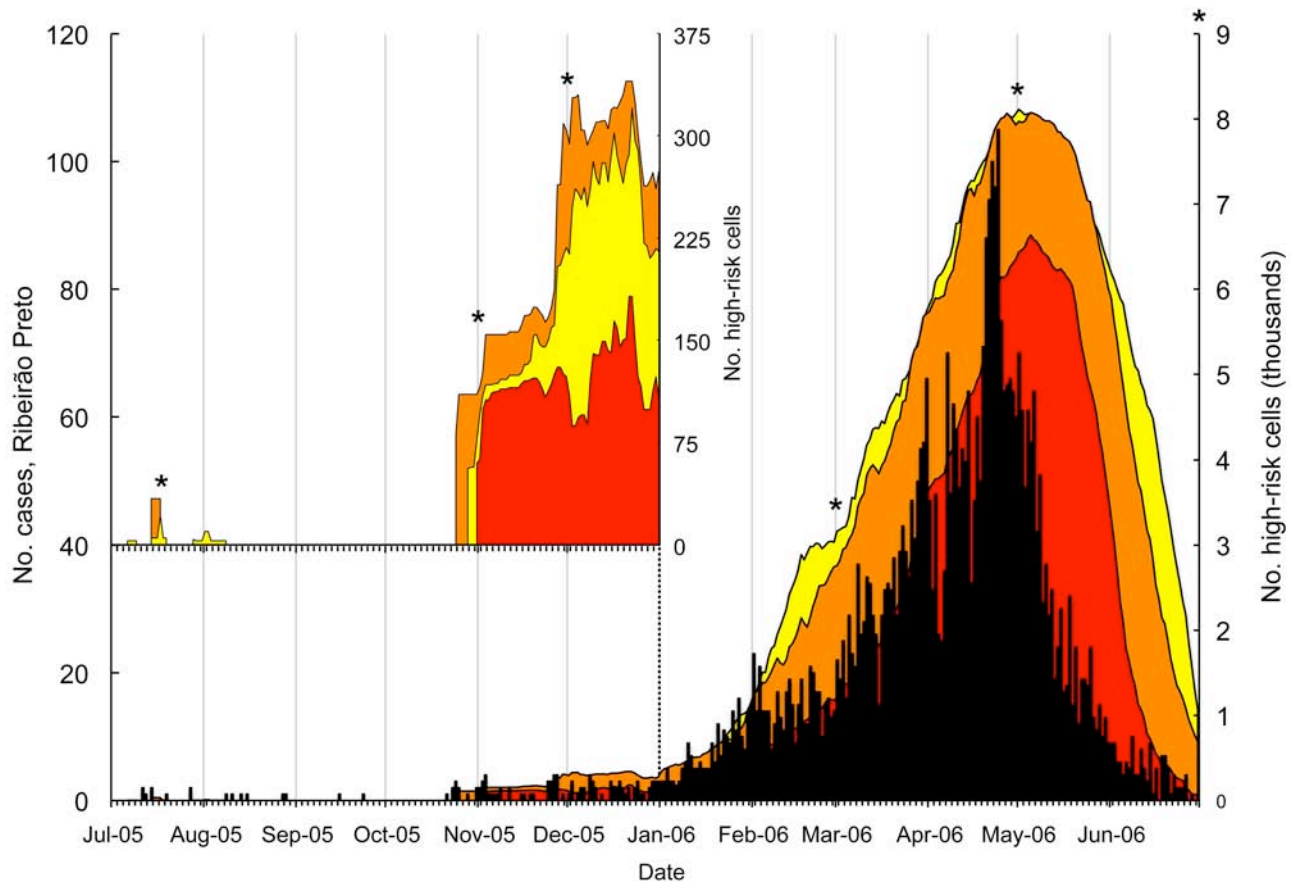


Figure 10. Number of confirmed dengue cases and DYCAST cells identified as high risk by models D10 (red), D05 (orange), and D10' (yellow) per day, Ribeirão Preto, Jul 1, 2005–Jun 30, 2006. Inset x-axis has the same scale, Jul 1, 2005–Jan 1, 2006; order of models D05 and D10' have been reversed in the inset for visualization purposes. Asterisks denote dates at which risk maps are shown in Figure 11.

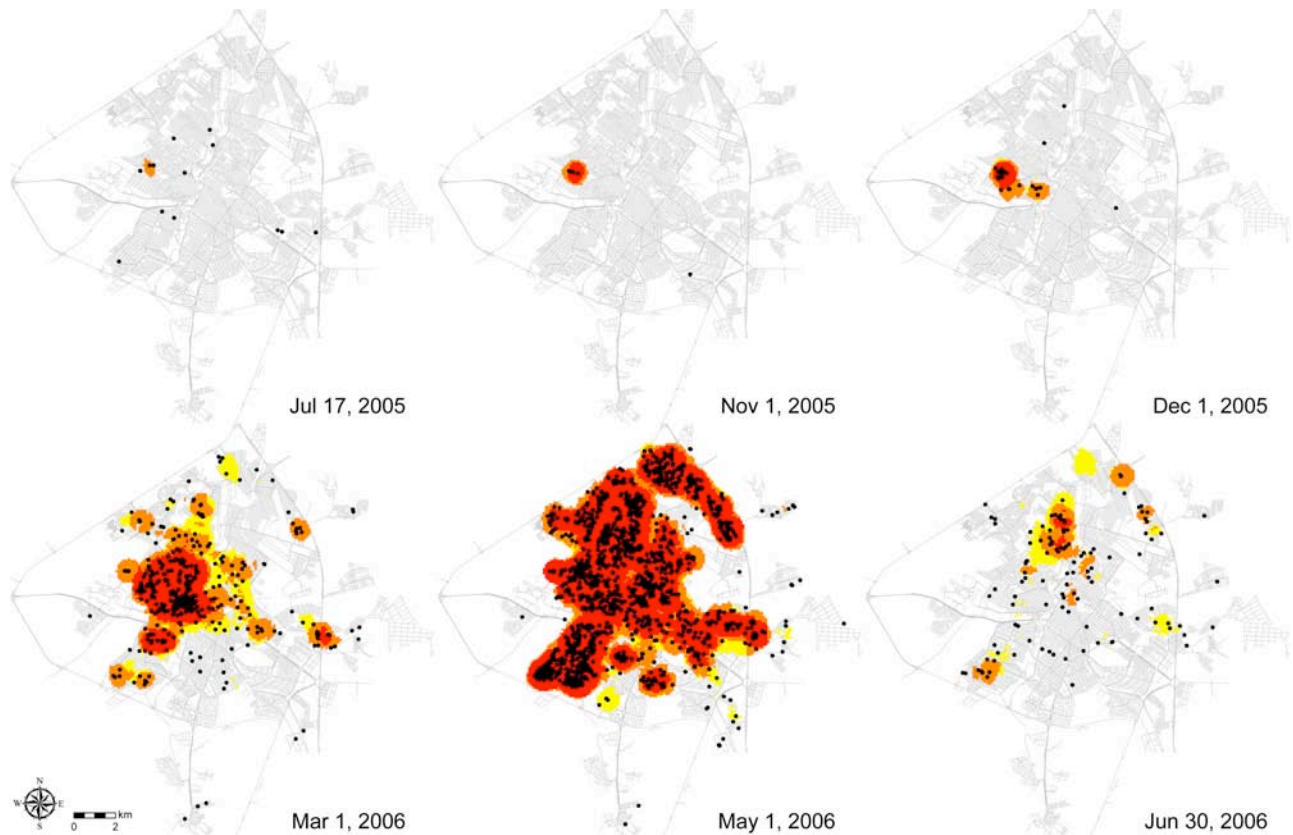


Figure 11. Confirmed dengue cases and DYCAST risk maps displaying high-risk cells as identified by models D10 (red), D05 (orange), and D10' (yellow), Ribeirão Preto. Dates correspond to asterisks in Figure 10. Maps illustrate presumed cryptic interepidemic focus of transmission (Jul 17) as well as its reemergence (Nov 1) and spread to surrounding areas (Dec 1), followed by increasing (Mar 1), peak (May 1), and contracting (Jun 30) levels of transmission, as defined by the number of high-risk cells. Cases displayed are those with an onset of illness on or up to 28 days prior to given date; this corresponds to the temporal domain of the model and thus includes all cases contributing to risk at that time. (see Supplemental Files for an animation of D10' risk maps during entire study period).

## Discussion

### Summary of findings

Because there is no drug prophylaxis, human vaccine, or treatment available for DENV, mosquito control and personal protection measures are the only options available for reducing human morbidity and mortality; thus, early warning of high-risk areas would allow such efforts to be targeted in a timely and effective manner. Results from this study, which represents the first spatially explicit model that uses human cases to predict future dengue risk, indicate that the modified DYCAST model provided early and accurate identification of high risk areas in Ribeirão Preto, including detection of what appears to be the cryptic interepidemic focus of transmission that later developed into the severe 2006 epidemic. Additionally, model efficacy was retained and even enhanced by including unconfirmed dengue cases in the analysis, which has important implications for increasing the utility and applicability of the model. These findings suggests that the DYCAST system could be utilized prospectively and in real-time to identify areas at high risk of DENV transmission, in order to target mosquito control, surveillance, and public education campaigns in a timely, efficient, and cost-effective manner. Furthermore, and in a departure from previous studies, this risk model was implemented using free, open-source, and cross-platform software that could provide an inexpensive and scalable GIS solution for the surveillance and control of DENV – and potentially other infectious diseases – by Ribeirão Preto and other public health agencies in the future.

### **Significance of findings, assessment of applicability to current theory and practice**

The DYCAST model's average sensitivity (across all specificities) of 78.3% is given by the AUC (0.783) (42). However, this should be considered an approximation, as only three data points were used to plot the ROC; the D10' model used different inputs than the other models as well, albeit its inclusion resulted in only a negligible increase in AUC (0.003). Youden's J statistic, the sum of each models' sensitivity and specificity over that of random chance, ranges from 0 (random chance) to 1.0 (perfect agreement; also -1.0 for perfect disagreement) and is useful if sensitivity and specificity are of equal importance in determining an optimal threshold (43). Model D10' had the highest J index as well as sensitivity, and would therefore be considered the optimal model; only if specificity were of primary concern would model D10 be considered optimal.

Sensitivity values and prediction rates are interpreted to be quite high, given that the dengue models use cases to predict future cases, and as such the incipient cases will inherently reduce both measures by never being predicted (albeit this may have been partially offset by considering as predicted any cells identified as high-risk on the date of onset of its earliest case). Furthermore, both the sensitivity and prediction rate of models D10' and D05 are greater than that of the WNV DYCAST model implemented during the unprecedented 2005 epidemic in California. This model, which used reports of dead birds to predict human cases, had a sensitivity of 80.8% (269/333 cells) and a prediction rate of 81.6% (289/354 cases) (19). However, specificity (90.6%; 66,543/73,434) was higher than that of all three dengue models.

These differences between the dengue and WNV models may reflect dissimilar viral ecologies. As a zoonotic disease, WNV transmission may be curtailed through die-

off and migration of infected and susceptible avian hosts, dynamics that do not apply to the DENV cycle. Rather, continuous presence of dense human populations, coupled with much smaller flight range and domestic proclivities of the *Ae. aegypti* vector, may account for more sustained viral transmission at a smaller scale; indeed, cells were identified as high-risk for a much longer total mean duration by the dengue models (65.4–92.3 days) than by the WNV model (39.0 days) (19). This phenomenon may also account in part for the observation that during the study period, half of all cases (50.1%; 2,352/4,690) occurred in cells containing a prior case; 44.4% of cells (1,039/2,338) that contained at least one case contained multiple cases (mean 2.0, maximum 31). Of course, these findings may also be explained by the fact that a single mosquito can infect multiple persons within the same residence.

Regardless, this redundancy of cases within single cells accounts for the differences between prediction rates (percentage of cases predicted) and model sensitivity (percentage of cells that predicted their earliest case out of all cells that contained cases); these differences equated to 12.1%, 7.2% and 6.5% for models D10, D05, and D10', respectively. Additionally, this redundancy allows for prediction rates to be compared against a simple null “model,” whereby cells are identified as high risk once a case occurs within them. This is equivalent to the aforementioned percentage of cases that occurred in cells containing a prior case, which provides a prediction baseline of 50.1% (2,352/4,690 cases). Models D10, D05, and D10' were thus  $\approx 1.6$ – $1.8x$  better at predicting cases than this null model, based on an additional 1,358 (+29.0%), 1,882 (+40.1%), and 1,885 (+40.2%) cases predicted, respectively.



Results from all three models also indicated relatively high accuracy (percent agreement), as well as a moderate strength of chance-adjusted agreement ( $0.40 < \kappa < 0.60$ , 44); kappa can be interpreted as the percentage of correctly identified cells expected to be misidentified by chance alone. PPV and NPV are a function of the prevalence of the outcome of interest, thus the low prevalence of cells that contained a human case (9.2%; 2,338/25,487) resulted in relatively low PPV and high NPV. Because RR is not influenced by prevalence, it is a more useful (albeit proxy) measure in this circumstance for communicating model predictiveness. This measure indicates that the RR of DENV infection was >10x higher in cells identified as high-risk compared to those not identified as such.

One of the most surprising findings in this study was that running the model with both confirmed and unconfirmed cases provided earlier and greater prediction of human cases, as well as higher sensitivity and RR, than using confirmed cases alone. These additional records represented an increase of 95.9% (4,496/4,690 records) over the confirmed cases, and consisted of seemingly useless information – 96.4% (4,335/4,496) of unconfirmed cases were diagnosed as or tested negative for DENV. However, the efficacy (and superiority) of model D10' may be explained by the following two factors. First, the model's statistical procedure for identifying significant clustering of cases in space and time should effectively eliminate the random noise in the data, and the fact that the procedure is based on dengue-specific parameters should reduce detection of significant clustering due to non-dengue illnesses. Secondly, there may actually be an epidemiologic signal of dengue in the non-confirmed case data. This may be due to the presence of undetected dengue cases within records that were missing data (n=159) or

due to false clinical diagnoses or laboratory findings that were negative (n=4,335) or inconclusive (n=2) for dengue. This is supported by the fact that unconfirmed cases generally follow annual trends of dengue incidence (Figure 12), however this may primarily be an artifact of greater vigilance and medical attention-seeking behavior by the population during dengue epidemics.

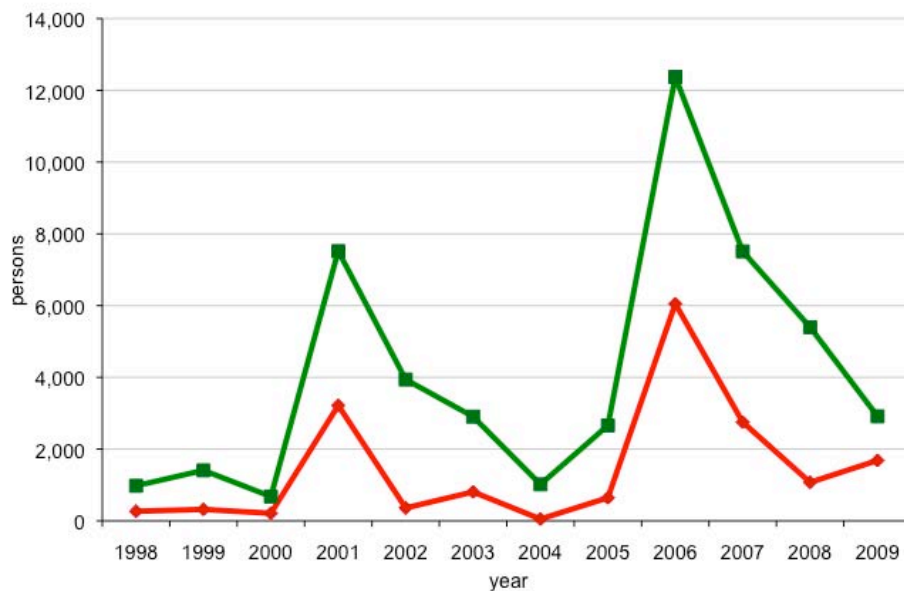


Figure 12. Number of suspected cases confirmed (red) and unconfirmed (green) for dengue infection per year, 1998-2009, Ribeirão Preto.

However, it is important to note that the increase in sensitivity afforded by D10<sup>7</sup> comes at a 6.5% reduction in specificity over D10. Presumably, this is due to the contribution of non-dengue cases to the detection of artificially significant clustering events. One solution would be to run multiple models at the same time (Figure 11), which would maximize both sensitivity and specificity. Alternatively, each model could be run

at different times of the year when different indices are preferred, such as running the more sensitive D10' during the interepidemic period and running the more specific D10 during the epidemic period. Furthermore, because D05 also provides better prediction and sensitivity than D10, it may be possible that a further increase in sensitivity could be achieved by creating a hybrid of D10' and D05 that would analyze all suspected cases at a threshold of 5.

From a practical standpoint, these results suggest that the model may retain efficacy in areas where thorough testing is not available or practical due to lack of resources or infrastructure; this has implications for making DYCAST applicable to a wider audience, as well as for the utility of modeling syndromic surveillance data. Additionally, the ability to predict high-risk areas without having to confirm suspected dengue cases eliminates the delay incurred from laboratory testing and clinical diagnoses, providing more timely results and thus a more rapid response to disease transmission.

Even without such an advantage of earlier case reporting, model D10' predicted cases with a greater mean number of days prior to onset than did the other models. Regardless, all models predicted cases with a mean length of time prior to onset sufficient for mosquito control efforts to respond and potentially interrupt viral transmission, even accounting for the 6-day human incubation period. The models' ability to provide early warning of disease transmission appears not to be a simple consequence of the large number of multiple cases in single cells (which increases the average duration); based only on the earliest cases within cells, the mean number of days that a cell successfully predicted its earliest case was 55.7 (std dev 38.0, median 51, max 292), 51.4 (std dev

38.5, median 46, max 295), 31.7 (std dev 31.7, median 32, max 201) for models D10', D05, and D10, respectively.

Another important finding is that as early as July 2005, DYCAST identified what appears to be the primary interepidemic focus of transmission that preceded the severe 2006 epidemic. This suggests the possibility that high-risk areas identified during the interepidemic period may reflect overwintering of the virus. Another notable feature of this detection is that without the risk maps, the clustering of cases is not conspicuous, and in fact appears against a backdrop of other seemingly scattered cases (Figure 11, Jul 17, 2005). This indicates that the model's statistical approach may add value to current surveillance and control efforts. Ultimately, such detection of otherwise cryptic transmission foci provides an opportunity for targeted mosquito control efforts to interrupt viral amplification before it reaches "critical mass" and spills over to surrounding areas.

### **Limitations of study and findings**

One limitation of the study includes potential selection bias of dengue cases, in that addresses not included in the street reference data, such as those in the rural periphery of the town or in newly developed areas, would not be geocoded. This may affect the case incidence rate as well as the sensitivity of the risk model. Future work could identify the presence of this bias through assessment of spatial structuring of geocoding rates. Additional bias could result from a patients' working address being recorded instead of their residential address. While difficult to assess posteriorly, this and

similar errors could be mitigated through quality control and manual rematching of unmatched addresses by someone local and familiar with the area.

A related bias concerns one of the primary assumptions of the model, which is that illness is acquired at a person's place of residence. While this is particularly consistent with the domestic nature of DENV transmission, illnesses acquired outside of a person's residence may confound the results. The second assumption of the model, that DENV transmission is spatially continuous, may be violated by virtue of the fact that persons can also transmit the virus to mosquitoes outside of their place of residence as well. Provided sufficient rates of vertical transmission, this assumption may also be violated by accidental human transport of *Ae. aegypti* eggs and larvae, which is reported in Brazil to be more important than adult flight for dispersion of the vector (29). Nevertheless, given the spatially confined flight range of the adult form that actually transmits the virus, violations of this assumption are considered to be negligible.

The third assumption of the model is that non-random spatiotemporal clustering indicates amplification of DENV. While this assumption is fully met by using only confirmed dengue cases, it may be violated by including non-dengue cases, as likely illustrated by the 5.9% reduction in specificity of D10' compared to that of D10. However, as mentioned above, the model's dengue-specific statistical procedure should generally limit the influence of such violations. Fourth, it is assumed that each dengue case has an equal opportunity of being reported. Differential rates of symptomatic DENV illness among various immunologic, age, or demographic groups may violate this assumption if such populations exhibit sufficient spatial heterogeneity (e.g., distribution of DENV antibodies within the population following spatially-structured epidemics).

This and the fact that DENV is asymptomatic in the majority of infections (45) may also yield false positive identifications, thereby reducing validity measures such as specificity and PPV. Unfortunately, these and other reporting biases are difficult to rigorously assess in the absence of serologic surveillance information. Symptomatology considerations aside, however, the existence of free universal healthcare in Brazil and prevalent media campaigns informing citizens about DENV symptoms may ensure that this assumption is generally met.

Other limitations concern the implementation of the software itself. The tradeoff with using multiple open source software components is that compatibility issues may arise as new versions of the various components are released. This was found to be the case in this study, and prompted the creation of protocols and a troubleshooting section to facilitate future implementations of the model (see Appendix). Additionally, running the model is computationally expensive, and required between 30 seconds to 2.5 hours to complete risk analysis for a single day. Thus, a dedicated computer with powerful hardware is recommended.

### **Relevant recommendations**

Future work will include running the dengue risk model(s) for additional years, to assess whether findings are similar during more (i.e., 2010) or less severe epidemics. One important avenue for future research includes the biology of *Ae. aegypti* from Ribeirão Preto populations, as findings from previous studies are often inconsistent and/or location-specific. Three main research objectives for informing the DYCAST model parameters are recommended. First, mark-release-recapture (MRR) studies in Ribeirão

Preto could determine the mean and maximum dispersion of the vector, which would elucidate the optimal spatial domain and spatial closeness parameters for this study area. Second, research should examine the role of temperature in affecting the lengths of the EIP and gonotrophic cycle of *Ae. aegypti*, which would inform the temporal domain and temporal closeness parameters. Such findings could be incorporated by specifying different parameters throughout the year, such as longer temporal parameters for a winter (interepidemic) model and shorter durations for a summer (epidemic) model, or by adjusting the relevant parameters based on real-time temperature data.

Additionally, population census data could be sampled at each grid cell's centroid at a radius equal to the model's 600 m spatial domain, in order to generate Monte Carlo distributions that reflect the spatial heterogeneities of the underlying human population (46). A more sophisticated approach, dasymetric mapping (47), could be used to re-aggregate population counts to only populated regions within a given census tract, as defined by remote sensing-derived land use classification imagery and/or residential lots data. This would yield a more accurate population layer by reducing error from the ecological fallacy of choropleth mapping.

Recommendations for future program development include the establishment of a pilot program to create an open source GIS network for electronic mapping and sharing of otherwise fragmented datasets used for surveillance and control by the various local health departments, vector control agency, and municipal government (CODERP) in Ribeirão Preto. This would provide for routine and timely mapping of surveillance data, and would allow for more rapid, coordinated, and efficient management across all facets of disease control, from public education campaigns and clinical surveillance to

prospective modeling and targeted vector control. This integrated framework would also better facilitate the practical application of the current research findings and act as a catalyst for further in-depth research.

This program would also serve as a replicable template for deploying scalable, no-cost GIS solutions that can be adopted by resource-limited areas around the world for the surveillance and control of dengue and other infectious diseases. Rather than relying on expensive commercial software packages, this project could leverage the free and open source software solutions used in the present study, thus providing a cheap but powerful GIS network with potential for web and mobile deployment. Future surveillance data could be collected by local agencies and entered in electronic format via PostgreSQL clients or online interfaces, created by using the Python Web Processing Service (PyWPS), for access remotely or via mobile devices; dynamic web content, such as case and risk maps used for public education campaigns, could be published by using Hypertext Preprocessor (PHP) scripting language. Security could be maintained by using Secure Sockets Layer (SSL) protocol to encrypt client/server communications, as well as password protection of computer accounts. Case data entered from the backend could be automatically geocoded and mapped; support for Geospatial Data Abstraction Library (GDAL) via the GDAL-GRASS plug-in would enable integration with other software as needed. Quantum GIS (QGIS; [www.qgis.org](http://www.qgis.org)) could be implemented for a more user-friendly frontend that includes a GUI in Portuguese, or GRASS GIS, which supports various visualizations (2D, 2.5D, and 3D raster and vector data) and spatial analyses. The latter could include R ([www.r-project.org](http://www.r-project.org)), coupled with the *crimestats*, *spatstat*, and *splan* packages for additional spatial and spatiotemporal analyses.



Regardless of these additional features, installation on a single computer of the components used in this study would allow for the prospective implementation of the dengue DYCAST model, which could generate daily risk maps in real-time for the immediate use in surveillance and control efforts. Protocols are also provided herein for installing the various software components, modifying the Python scripts to model dengue risk, and generating the necessary spatial datasets (see Appendix). The open source version of the DYCAST software and documentation (see Acknowledgments) can be downloaded at [www.dycast.org](http://www.dycast.org), a web domain purchased by the present study's author to facilitate the dissemination of the model and results to a global audience, and which will be fully developed in the future. Ultimately, the practical application of these modeling efforts via a GIS pilot program may act as a catalyst for further research and open source collaboration, and enhance public health infrastructure of the developing world that bears the greatest burden of dengue morbidity and mortality.

## H. References

1. Gubler DJ. 2002. Epidemic dengue/dengue hemorrhagic fever as a public health, social and economic problem in the 21st century. *Trends Microbiol.* 10:100-3.
2. World Health Organization. 2006. Report of the Scientific Working Group on Dengue. Geneva, 1-5 October 2006. 168 pages.  
<http://apps.who.int/tdr/svc/publications/tdr-research-publications/swg-report-dengue>
3. World Health Organization. 1999. Strengthening implementation of the global strategy for Dengue fever and Dengue haemorrhagic fever, prevention and control. Report in the informal consultation. Geneva: WHO HQ. 20 pages.
4. Nimmannitya S. 1997. Dengue hemorrhagic fever: diagnosis and management. Gubler DJ, Kuno G, eds. *Dengue and dengue hemorrhagic fever*. New York: CAB International. pp 133-46.
5. Innis BL. 1995. Dengue and dengue hemorrhagic fever, In J. S. Porterfield (ed.), *Exotic viral infections – 1995*. Chapman & Hall, London, United Kingdom. pp 103-46.
6. Hay SI, Myers MF, Burke DS, Vaughn DW, Endy T, Ananda N, Shanks GD, Snow RW, Rogers DJ. 2000. Etiology of interepidemic periods of mosquito-borne disease. *Proc Natl Acad Sci USA* 97(16):9335-9.
7. Passos ADC, Rodrigues EMS, Dal-Fabbro AL. 1998. Dengue control in Ribeirão Preto, São Paulo, Brazil. *Cad. Saúde Pública*, Rio de Janeiro 14(Suppl. 2):123-8.

8. Figueiredo LT, Owa MA, Carlucci RH, dal Fabbro AL, de Mello NV, Capuano DM, Santili MB. 1995. Dengue serologic survey in Ribeirão Preto, São Paulo, Brazil. *Bull Pan Am Health Organ* 29(1):59-69.
9. Ribeirão Preto Online. Ribeirão Preto já soma 31 casos de dengue hemorrágica (Ribeirão Preto already has 31 cases of dengue hemorrhagic [fever]). Updated 2010 Apr 23. Ribeirão Preto, Brazil; 2010 [cited April 28, 2010]. Available from: <http://www.ribeiraopretoonline.com.br/saude/ribeirao-preto-ja-soma-31-casos-de-dengue-hemorragica/35308>
10. Scott TW, Morrison AC. Vector dynamics and transmission of dengue virus: implications for dengue surveillance and prevention strategies. In: A.L. Rothman (ed.), *Dengue Virus, Current Topics in Microbiology and Immunology* 338:115–28.
11. Focks DA, Daniels E, Haile DG, Keesling JE (1995) A simulation model of the epidemiology of urban dengue fever: literature analysis, model development, preliminary validation, and samples of simulation results. *Am J Trop Med Hyg* 53:489–506.
12. Luz PM, Codeço CT, Massad E, Struchiner CJ. 2003. Uncertainties Regarding Dengue Modeling in Rio de Janeiro, Brazil. *Mem Inst Oswaldo Cruz* 98(7):871–8.
13. Honório NA, Nogueira RMR, Codeço CT, Carvalho MS, Cruz OG, et al. 2009 Spatial evaluation and modeling of dengue seroprevalence and vector density in Rio de Janeiro, Brazil. *PLoS Negl Trop Dis* 3(11):e545.  
doi:10.1371/journal.pntd.0000545

14. Barreto FR, Teixeira MG, Costa MCN, Carvalho MS, Barreto ML. 2008. Spread pattern of the first dengue epidemic in the city of Salvador, Brazil. *BMC Public Health* 8:51–71.
15. Siqueira JB, Martelli CM, Maciel IJ, Oliveira RM, Ribeiro MG, Amorim FP, Moreira BC, Cardoso DD, Souza WV, Andrade AL. 2004. Household survey of dengue infection in central Brazil: spatial point pattern analysis and risk factors assessment. *Am J Trop Med Hyg* 71(5):646–51.
16. Morrison AC, Getis A, Santiago M, Rigua Perez JG, Reiter P (1998) Exploratory space time analysis of reported dengue cases during an outbreak in Florida, Puerto Rico, 1991–1992. *Am J Trop Med Hyg* 58:287–98.
17. Theophilides CN, Ahearn SC, Grady S, Merlino M. Identifying West Nile virus risk areas: the Dynamic Continuous-Area Space-Time system. *Am J Epidemiol.* 2003;157(9):843–54.
18. Theophilides, C. N., et al. First evidence of West Nile virus amplification and relationship to human infections. *International Journal of Geographical Information Science* 2006;20(1):103–15.
19. Carney RM, Ahearn SC, McConchie A, Glaser C, Jean C, Barker C, Park B, Padgett K, Kramer V. Early warning system for West Nile virus risk areas: the Dynamic Continuous-Area Space-Time system, California. *Emerg Infect Dis.* *in peer review* [as of 05-01-2010]
20. California Department of Public Health. Mosquito and Vector Control Association of California; University of California. California Mosquito-borne Virus Surveillance & Response Plan. Sacramento (CA): The Department. Vector-

- Borne Disease Section. 2009 Apr [cited 2010 Apr 29]. Available from <http://www.cdph.ca.gov/HealthInfo/discond/Documents/2009MosqSurvRespPlan.pdf>
21. Carney RM, Husted S, Jean C, Glaser C, Kramer V. Efficacy of aerial spraying of mosquito adulticide in reducing incidence of West Nile virus, California, 2005. *Emerg Infect Dis*. 2008;14:747–54. DOI: 10.3201/eid1405.071347
  22. Knox EG. Detection of low intensity epidemicity: application to cleft lip and palate. *Brit J Prev Soc Med*. 1963;17:121–7.
  23. Knox EG, Bartlett MS. The detection of time-space interactions. *Appl Stat*. 1964;13(1):25–30.
  24. Harrington LC, Scott TW, Lerdthusnee K, Coleman RC, Costero A, Clark GG, Jones JJ, Kitthawee S, Kittayapong P, Sithiprasasna R, Edman JD. 2005. Dispersal of the dengue vector *Aedes aegypti* within and between rural communities. *Am J Trop Med Hyg* 72:209–20.
  25. Hausermann W, Fay RW, Hacker CS. 1971. Dispersal of genetically marked female *Aedes aegypti* in Mississippi. *Mosquito News* 31(1):37–51.
  26. Maciel-De-Freitas R, Codeço CT, Lourenço-De-Oliveira R. 2007. Body size-associated survival and dispersal rates of *Aedes aegypti* in Rio de Janeiro. *Medical and Veterinary Entomology* 21:284–2.
  27. Honório NA, Silva WDC, Leite PJ, Gonçalves JM, Lounibos LP, Lourenço-de-Oliveira R. 2003. Dispersal of *Aedes aegypti* and *Aedes albopictus* (Diptera: Culicidae) in an Urban endemic dengue area in the state of Rio de Janeiro, Brazil. *Mem Inst Oswaldo Cruz, Rio de Janeiro* 98(2):191–8.
  28. Gubler DJ. 1998, *Clin Microbiol Rev* 11(3):480–96.

29. National Health Foundation [Fundação Nacional de Saúde]. 2001. Dengue: instructions for vector control personnel – manual of technical standards [in Portuguese]. 3rd Ed., rev. Brasília: Ministério da Saúde. 84 pp.
30. Salazar MA, Richardson JH, Sánchez-Vargas I, Olson KE, Beaty BJ. 2007. Dengue virus type 2: replication and tropisms in orally infected *Aedes aegypti* mosquitoes. *BMC Microbiology* 7(9): DOI:10.1186/1471-2180-7-9
31. Siler JF. 1926. Dengue: its history, epidemiology, mechanism of transmission, etiology, clinical manifestations, immunity, and prevention. *Philippine J of Science* 29:170–210.
32. Simmons JS. 1931. Dengue fever. *Am J Trop Med Hyg* 11(2):77–82.
33. Christophers SR. 1960. *Aedes aegypti* (L). The yellow fever mosquito, its life history, bionomics and structure. Cambridge University Press, London. 739 pp.
34. Morlan HB, Hayes RO. 1958. Urban dispersal and activity of *Aedes aegypti*. *Mosquito News* 18(2):137–144.
35. McClelland GAH, Conway GR. 1971. Frequency of blood feeding in the mosquito *Aedes aegypti*. *Nature* 232:485–6.
36. Seawright JA, Dame DA, Weidhaas DE. 1977. Field survival and ovipositional characteristics of *Aedes aegypti* and their relation to population dynamics and control. *Mosquito News* 37(1):62–70.
37. Marchoux E, Salimbeni A, Simond PL. 1903. Yellow Fever [in French]. *Rapports de la Mission Française. Ann Inst Pasteur* 17:665–731.
38. Mantel N. The detection of disease clustering and a generalized regression approach. *Cancer Research*. 1967;27(2):209–20.
39. Szklo M and Nieto FJ. 2007. *Epidemiology: Beyond the Basics* (2nd Ed.)

40. Byrt T, et al. 1993. Bias, prevalence, and kappa. *J Clin Epidemiology*. 46(5):423-9.
41. Hanley JA, McNeil BJ. The meaning and use of the area under a ROC curve. *Radiology*. 1982;143:27–36.
42. Glas AS, Lijmer JG, Prins MH, Bossel GJ, Bossuyt PMM. 2003. The diagnostic odds ratio: a single indicator of test performance. *J Clin Epi* 56(2003):1129–1135
43. Youden WJ. Index for rating diagnostic tests. *Cancer*. 1950;3:32-5.
44. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159–74.
45. Gubler DJ and Kuno G (ed). 1997. *Dengue and dengue hemorrhagic fever*. CAB International, London, United Kingdom.
46. Kulldorff M, Hjalmar U. 1999. The Knox method and other tests for space-time interaction. *Biometrics* 55(2):544–52.
47. Langford M. and D. J. Unwin. 1994. Generating and mapping population density surface within a geographical information system. *The Cartographic Journal* 31:21–6.

---

## Appendix I. Protocol – software installation, modification, FAQ

---

### Installation (Windows):

1. **PostgreSQL** <http://www.postgresql.org/download/windows>
  - install **PostgreSQL 8.4.3-1** (one-click installer, not the pgInstaller)
  
2. **PostGIS**
  - install **PostGIS v 1.5.1** using the PostgreSQL installer's Stack Builder, since it offers the latest version - no need to install separately afterwards. (latest version: <http://postgis.refractor.net/download/windows/>)
  - enable the **shp2pgsql** graphical loader plugin (will use later for grid creation)
  
3. **Python** <http://www.python.org/download/releases/2.6.5/>
  - module pycogp2 (below) requires Python 2.6.x and earlier [see **Troubleshooting** below for conflict with ArcView]
  - install Windows **64-bit version\* of Python 2.6.5**
  - restart computer
  
4. **Psycogp2** <http://www.stickpeople.com/projects/python/win-psycogp/>
  - if applicable, install the **64-bit version\* of psycogp 2.0.14 “(For Python 2.6 amd64)(64bit Windows)”**; be sure to use the specific build of psycogp2 for your specific version of Python.
  - restart computer

*\* “The binaries for AMD64 will also work on processors that implement the Intel 64 architecture (formerly EM64T), i.e. the architecture that Microsoft calls x64, and AMD called x86-64 before calling it AMD64. They will not work on Intel Itanium Processors (formerly IA-64).”*



## 5. DYCAST

- From [www.dycast.org](http://www.dycast.org), download and unzip **application.zip**. In the application folder:
- delete the `psycopg2` folder in `DYCAST/application/lib`
- Open **dycast.config** file in WordPad or Notepad and make the following changes:

1. under [database]:

change password from: **postgres**  
to: [ insert PostgreSQL password ]

2. under [dycast]:

spatial domain: 0.372822 [600 meters]  
close\_in\_space: 0.062137 [100 meters]  
close\_in\_time: 4  
temporal\_domain: 28  
bird\_threshold: 10 [analysis threshold]

[Note: need to close and restart **ui.py** every time you change parameters]

3. under [other]:

spatial\_reference\_unprojected: 29193  
spatial\_reference\_projected: 29193

- Save and close

3. Right-click **postgres\_init.sql** (also in application folder) > Edit (opens in Notepad)

- do Find-Replace All to change all instances of 4269 and 54003 to **29193**  
This is the EPSM code relevant for Ribeirão Preto. Save and close

4. Right-click **dycast.py** > Edit with IDLE

- line 200: change: 4269 and 54003  
to: **29193**

- line 493 [“def get\_county\_id(tile\_id):” statement]:

change this:                    **return cur.fetchone()[0]**  
to this:                            **return 1**

- Save and close

5. Right-click **setup\_db.bat** > Edit

- PostgreSQL 8.4 installs into a different directory than 8.2, and also has a different share path. Therefore, change lines 12 and 13 to:

```
set PG_BIN_PATH=C:\Program Files (x86)\PostgreSQL\8.4\bin\  
set PG_SHARE_PATH=C:\Program Files  
(x86)\PostgreSQL\8.4\share\contrib\postgis-1.5\  

```

- Another change in 8.4 is that the file **lwpostgis.sql** is renamed to **postgis.sql**. Therefore, change line 24 to:

```
%PG_BIN_PATH%\psql.exe" -U %USERNAME% -d  
%DBNAME% -f "%PG_SHARE_PATH%\postgis.sql
```

- Save and close

6. Double-click **setup\_db.bat** to initialize database. Will prompt for password multiple times.

Installation (MacOS X):

1. **PostgreSQL 8.4.3-1** <http://www.enterprisedb.com/products/pgdownload.do#osx>

- Note: installer will automatically change shared memory settings, which may destabilize system (see Troubleshooting, below). Additionally, instructions for manually doing this are included in the Readme.

2. **PostGIS**

- the Stack Builder app that PostgreSQL runs offers PostGIS 1.4. The newest version (<http://postgis.refrains.net/download/>) is **1.5.0-1**, but this won't install (keeps saying Postgres 8.4 must be installed, which it is); this may have something to do with the location requirement [from ReadMe]: "Postgres must be installed in the default location (/usr/local/pgsql), as is my Postgres package."
- Therefore, use PostgreSQL's Stack Builder to install PostGIS version 1.4.

3. **Python:**

- comes pre-installed on Mac OS X: Terminal > python:  
I have: **Python 2.6.1** (r261:67515, Jul 7 2009, 23:51:51)

4. **Psycopg2** <http://initd.org/psycopg/download/>

- download psycopg2 2.0.14
- in **setup.cfg** file, change line 31:

```
from          #pg_config=  
to:           pg_config=Library/PostgreSQL/8.4/bin/pg_config
```

- need **GNU C Compiler** to build psycopg2. Download gcc-4.2:  
<http://gcc.gnu.org/gcc-4.2/>

5. **DYCAST**

- put folder in: **Users/Ryan/Documents**
- follow **Steps 5.1 – 5.4** listed above (Windows installation) for modifying scripts.

- Also, modify **dycast.config** (open in TextEdit)

**[system]**

change:        unix\_dycast\_path:    /Users/alan/Documents/DYCAST/  
to:            unix\_dycast\_path:    /Users/Ryan/Documents/DYCAST/

**[database]**

change:        password:        postgres  
to:            password:                          

- modify **setup\_db.sh** (open in TextEdit)

**line 6**

change:        DYCAST\_PATH=/Users/alan/Documents/DYCAST/  
to:            DYCAST\_PATH=/Users/Ryan/Documents/DYCAST/

**line 9**

change:        #PG\_BIN\_PATH=/usr/local/pgsql/bin...?  
to:            PG\_BIN\_PATH=/Library/PostgreSQL/8.4/bin/

**line 10**

change:        PG\_SHARE\_PATH=/usr/local/pgsql/share/  
to:            PG\_SHARE\_PATH=/Library/PostgreSQL/8.4/share/postgresql/contrib

**line 15**

change:        psql -U \$USERNAME -d \$DBNAME -f  
                  \$PG\_SHARE\_PATH/lwpostgis.sql  
to:            \$PG\_BIN\_PATH/psql -U \$USERNAME -d \$DBNAME -f  
                  \$PG\_SHARE\_PATH/postgis.sql

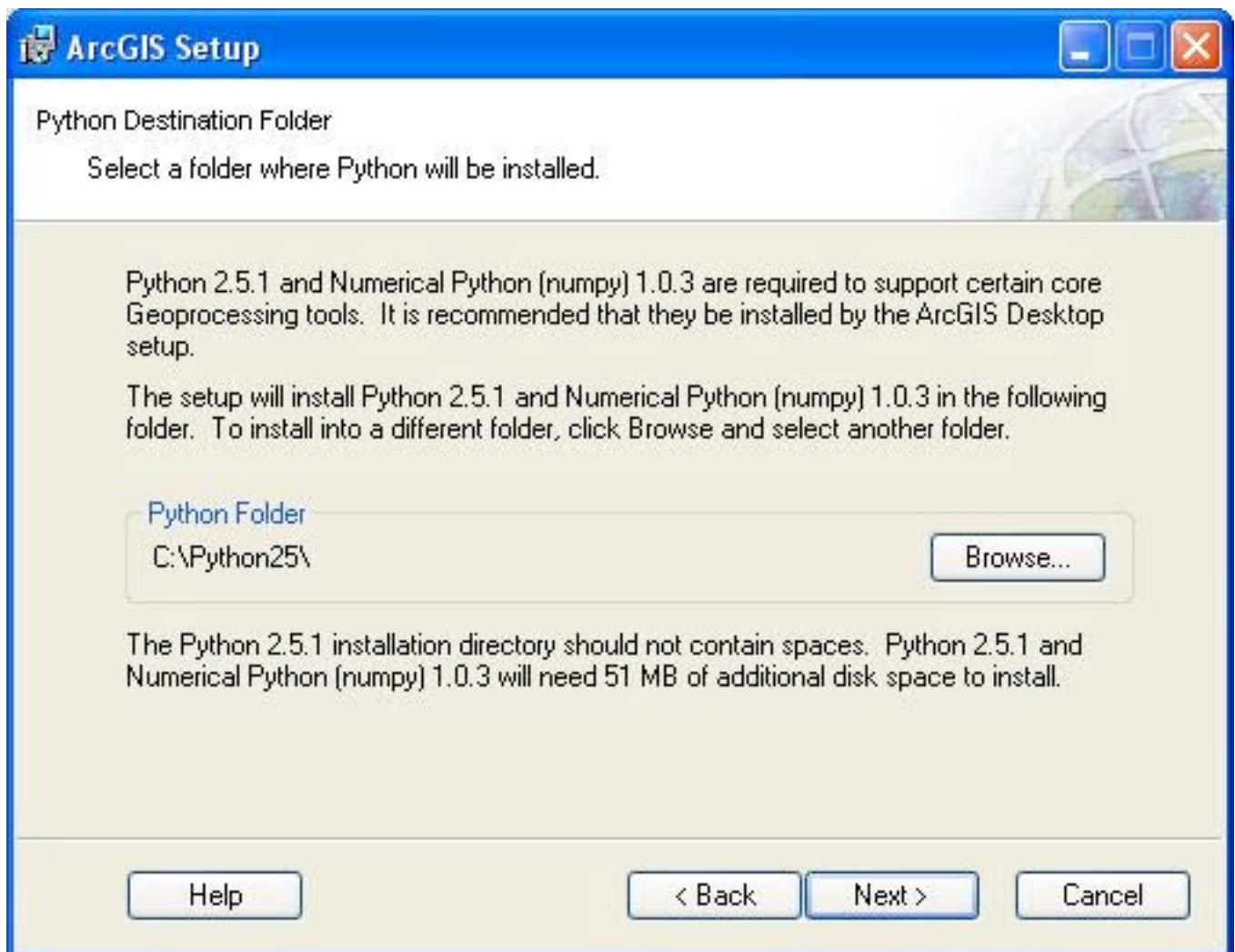
- lines 12-22: "\$PG\_BIN\_PATH/" needs to be inserted before each command (dropdb, createdb, createlang, etc.).
- open with (or copy-paste script into) **Terminal** to run setup
- Double-click **setup\_db** to initialize database. Will prompt for password multiple times.

Troubleshooting FAQ

1. **Problem:** [Windows]: DYCAST (file ui.py) no longer runs after installing ArcView 9.3.0:

```
couldn't import psychopg2 library in path: ['C:\\DYCAST\\application', 'C:\\Program Files (x86)\\ArcGIS\\bin', 'C:\\Windows\\system32\\python25.zip', 'C:\\Python25\\DLLs', 'C:\\Python25\\lib', 'C:\\Python25\\lib\\plat-win', 'C:\\Python25\\lib\\lib-tk', 'C:\\Python25', 'C:\\Python25\\lib\\site-packages', 'C:\\DYCAST\\application\\libs', 'libs', 'C:\\DYCAST\\application\\libs\\dbfpy', 'libs\\dbfpy', 'C:\\DYCAST\\application\\libs\\psychopg2', 'libs\\psychopg2']
```

**Cause:** ArcView installed an older version of Python (2.5.1, see below), which conflicts with the Python 2.6 already installed (psychopg2 looks for python in the Python25 folder – see above screenshot).



**Solution:** The only way to get around this incompatibility issue is to drag the ui.py file onto the copy of python in the **C:/Python26** folder. Otherwise, simply double-clicking on the file will have it try to run off the Python25 copy, which won't work.

### **Background Information:**

The full list of operations requiring Python/Numpy include (taken from "Python\_requirement.htm" file on installation DVD:

#### **Analysis Tools**

- Proximity Toolset
- Multiple Ring Buffer

#### **Conversion Tools**

##### **To dBASE**

- Table to dBASE (multiple)

##### **To Geodatabase**

- Feature Class to Geodatabase (multiple)
- Table to Geodatabase (multiple)

##### **To Shapefile**

- Feature Class to Shapefile (multiple)

#### **Spatial Statistics**

##### **Analyzing Patterns**

- Average Nearest Neighbor
- High/Low Clustering (Getis-Ord General G)
- Spatial Autocorrelation (Morans I)

##### **Mapping Cluster**

- Cluster and Outlier Analysis (Anselin Local Morans I)
- Hot Spot Analysis (Getis-Ord Gi\*)

##### **Measuring Geographic Distributions**

- Central Feature
- Directional Distribution (Standard Deviation Ellipse)
- Linear Directional Mean
- Mean Center
- Standard Distance

##### **Utilities**

- Calculate Areas
- Collect Events
- Count Rendering
- Export Feature Attribute to Ascii
- Z Score Rendering

2. **Problem:** [Mac OS]: Computer destabilizes after PostgreSQL installation, causing crashing, slowdown, and startup displaying blinking folder icon with “?” sign.

**Cause:** Conflicts due to reconfiguration of Shared Memory settings by PostgreSQL installer.

### Background Information:

PostgreSQL One Click Installer README

Shared Memory

PostgreSQL uses shared memory extensively for caching and inter-process communication. Unfortunately, the default configuration of Mac OS X does not allow suitable amounts of shared memory to be created to run the database server.

Before running the installation, please ensure that your system is configured to allow the use of larger amounts of shared memory. Note that this does not 'reserve' any memory so it is safe to configure much higher values than you might initially need. You can do this by editing the file `/etc/sysctl.conf` - e.g.

```
% sudo vi /etc/sysctl.conf
```

On a MacBook Pro with 2GB of RAM, the author's `sysctl.conf` contains:

```
kern.sysv.shmmax=1610612736
kern.sysv.shmall=393216
kern.sysv.shmmin=1
kern.sysv.shmmni=32
kern.sysv.shmseg=8
kern.maxprocperuid=512
kern.maxproc=2048
```

Note that  $(\text{kern.sysv.shmall} * 4096)$  should be greater than or equal to `kern.sysv.shmmax`. `kern.sysv.shmmax` must also be a multiple of 4096.

Once you have edited (or created) the file, reboot before continuing with the installation. If you wish to check the settings currently being used by the kernel, you can use the `sysctl` utility:

```
% sysctl -a
```

The database server can now be installed. For more information on PostgreSQL's use of shared memory, please see: <http://www.postgresql.org/docs/current/static/kernel-resources.html> - SYSVIPC

Support. For help with this installer, please visit the forum at:

<http://forums.enterprisedb.com/forums/show/9.page>

- When you run PostgreSQL installer, you receive this message:



...which modifies the following:

- my settings before (**Terminal** > `sysctl -a`):

```
kern.sysv.shmmax: 4194304
kern.sysv.shmmin: 1
kern.sysv.shmmni: 32
kern.sysv.shmseg: 8
kern.sysv.shmall: 1024
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 10
```

- my settings after PostgreSQL installation:

```
kern.sysv.shmmax: 33554432
kern.sysv.shmmin: 1
kern.sysv.shmmni: 256
kern.sysv.shmseg: 64
kern.sysv.shmall: 8192
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 87381
kern.sysv.semmani: 10
```



**Solution:** Running the uninstaller in Library/PostgreSQL will NOT automatically revert sysctl.conf settings, therefore you must do this manually:

1. Shut down computer
2. Restart and press **Command-S** while booting. This boots in single user mode.
3. At the prompt (“:/ root#”) type the following four commands in sequence:

```
/sbin/fsck -fy           [resulting procedure will take a moment]
/sbin/mount -uw /       [this mounts in read/write mode]
rm /etc/sysctl.conf
exit
```

Note that the last three commands will not provide any response, just reprompt you.

4. Restart computer.
5. Run **Terminal**, and type:

```
sysctl -a
```

Verify that kern.sysv settings have reverted to previous.

---

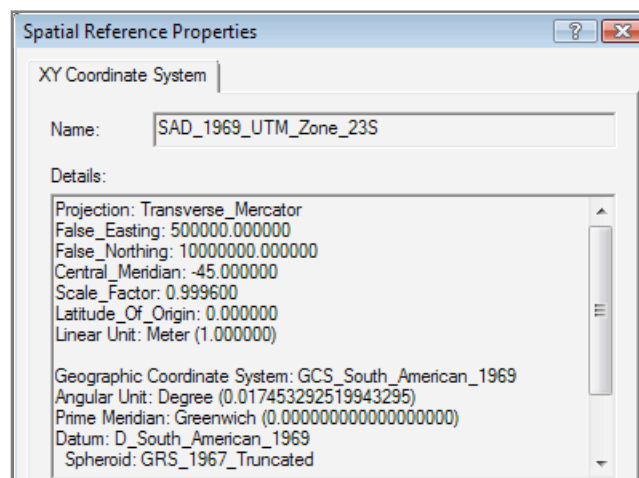
## Appendix II. Protocol – geocoding

---

### I. CREATE ARCMAP DOCUMENT

---

1. Open ArcMap and begin a new map document.
2. **Add aerial photograph (OPTIONAL)** [Note: this was taken 2005]
  - [update 2010-04-01] According to Wilson: “On the spatial reference of the photograph, all were designed using **North American Datum 1969 UTM Zone 23S.**” However, there is no NAD 1969, only NAD 1927 and NAD 1983, and they only have UTM Zones with “N”, not “S”. Perhaps he meant SAD 1969?
  - Open ArcCatalog and navigate to **Foto.TIF** > right-click > Properties > Spatial Reference [which should read “<Undefined>”] > Edit > Select > Projected Coordinate Systems > UTM > South America > **South American 1969 UTM Zone 23S.prj** \* > Add > OK:



Since there is no *original* projection file accompanying this raster file, the particular Spatial Reference (SAD\_1969\_UTM\_Zone\_23S) was determined from the following three pieces of evidence:

**A. Foto.tab file** (within the CODERP DVD folder) states:

```
!table
!version 300
!charset WindowsLatin1
```

```
Definition Table
File "300.TIF"
Type "RASTER"
(197457.283284349,7666274.3256114) (0,0) Label "Pt1",
```

```
(224235.096661752,7666274.3256114) (24991,0) Label "Pt2",  
(224235.096661752,7639495.44069285) (24991,24991) Label "Pt3"  
CoordSys Earth Projection 0, 0
```

This **CoordSys Earth Projection** means it is in some kind of UTM.

**B. CODERP DVD geodesic data use several UTM:** [see Monografias files]

Projection:	Internacional 1967	WGS-84	Internacional 1924 - Hayford
Datum:	SAD-69	WGS-84	Córrego Alegre

**C. ...IBGE map of Ribeirão Preto however, only uses:** [SP-RibeiraoPreto.pdf]

Projection:	UTM *
Datum (horizontal):	SAD 69

\* “origem da quilometragem UTM: Equador e Meridiano 45 W Gr. Acrescidas as constantes de 10.000 e 500 Km respectivamente” → “origin mileage **UTM Ecuador and Meridian 45 W Gr** Plus the constants of 10,000 and 500 km respectively.” The boldface denotes it is UTM 23S:

<http://home.pacbell.net/lgalvin/UTMWGS84CoordSys.html>

**D. Therefore, I tested both spatial references for Foto.TIF** [against RIBEIRAO\_PRETO\_SP with GCS\_South\_American\_1969 / D\_South\_American\_1969 (unprojected)]:

- **SAD\_1969\_UTM\_Zone\_23S:** [correct]



- WGS\_1984\_UTM\_Zone\_23S: [mismatched]



- Add **Foto.TIF** from CODERP DVD > *single-click* Foto.TIF to select (to retain all 3 bands) > Add. When prompted, clicked Yes to **build pyramids**.

### 3. Add streetlayer

Unzip **Arruamento.zip** file (Note: it contains the projection file). Add newly unzipped **RIBEIRAO\_PRETO\_SP.shp** from the Arruamento folder. Its spatial reference is the following\*:

GCS: **GCS\_South\_American\_1969**  
Datum: **D\_South\_American\_1969**  
Prime Meridian: Greenwich  
Angular Unit: Degree

\* Note: no mismatches exist between **RIBEIRAO\_PRETO\_SP.shp** and **Foto.TIF** layers when the Layers' data frame is either unprojected (a la former file's GCS) or projected (latter's **SAD\_1969\_UTM\_Zone\_23S**). However, the former's map units are in **Decimal Degrees**, which is why I used the latter (in **meters**) – this also provides perpendicular lat-long rasters/grids!

### 4. Project streetlayer

**ArcToolbox** > Data Management Tools > Projections and Transformations > Feature > **Project**

Input Dataset or Feature Class	<b>RIBEIRAO_PRETO_SP.shp</b>
Input Coordinate System (optional) [GCS_South_American_1969 should be displayed]	
Output Dataset or Feature Class	DATA\STREETS\ADDRESS_LOCATOR\ 03_2010_LOCATOR\Arruamento\ <b>streets_P_SAD69_UTM23S.shp</b>
Output Coordinate System	click button to right > Select > Projected Coordinate Systems > UTM > South America > <b>South American 1969 UTM Zone 23S.prj</b> > Add > OK

[will appear as **SAD\_1969\_UTM\_Zone\_23S**]

Geographic Transformation (optional) [null – none is needed] > OK

5. Resulting new shapefile should perfectly overlay the old, unprojected version. Remove the old one.

---

## II. OFFSET DETERMINATION

---

### 1. Create raster of distance from street values

- Spatial Analyst toolbar > Distance > **Straight Line**:  

Distance to:	<b>streets_P_SAD69_UTM23S</b>
Maximum distance:	[null]
Output cell size:	<b>1</b> [units in meters]
Create direction:	[leave box unchecked]
Create allocation:	[leave box unchecked]
Output raster:	<b>street_P_dist</b>

[right-click > Symbology tab > Classified > Classify button to manually re-classify symbology]

### 2. Add CODERP cases [n=11,909]

- Add Data > DATA\HUMANS\SHAPEFILES\I\_CODERP\_2007-2008\I\_ORIGINAL\_SAD69\_UTM23S\2008-06-11\_Dengue\_2007-2008\_cases\_SAD69\_UTM23S.shp

Ignore the following error message: “The following data sources you added are missing spatial reference information. This data can be drawn in ArcMap, but cannot be projected.”

[Note: spatial reference for shapefile cannot be added using ArcCatalog]

- **ArcToolbox** > Data Management Tools > Projections and Transformations > **Define Projection**

Input Dataset or Feature Class	<b>2008-06-11_Dengue_2007-2008_cases_SAD69_UTM23S</b>
Coordinate System	<b>South American 1969 UTM Zone 23S.prj</b> *

[Note: I subsequently projected the file as SAD69 as well to create the .PRJ file)

\* Note: I tested out two other UTM23S projections – **WGS84 (blue)** and **Corrego Alegre (yellow)** – and when I did an error stated, “Datum conflict between map and output.” **Red** = **SAD69**, which is also where the WGS and CA points were before defining their respective projections. Also reprojected WGS84 as SAD69 (not shown) but the result was even further off:



- Results: distance between SAD69 and WGS84 points  $\approx$ 8.2 meters
- **SAD69 appeared to have the better results** (see above, inset)
- Additionally, I did same test for CODERP's **vector area** and **health jurisdiction** shapefiles (both without .PRJ and spatial reference information), and *both* were unequivocally **SAD\_1969\_UTM\_Zone\_23S** as well.

### 3. Extract distance values at location of CODERP cases

- ArcToolbox > Spatial Analyst Tools > Extraction > **Extract Values to Points**

Input point features	2008-06-11_Dengue_2007-2008_cases_SAD69_UTM23S
Input raster	streets_P_dist
Output point features	2008-06-11_Dengue_2007-2008_cases_SAD69_UTM23S _dist_from_street.shp
Interpolate values...	[unchecked]

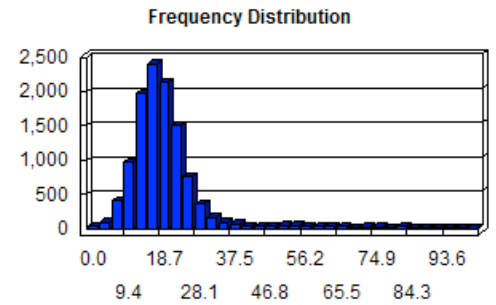
Append all the input... [unchecked]

• **RESULTS:**

	<u>ALL</u>	<u>&lt;100 m*</u>	<u>CLASSI_1-4**</u>	<u>CLASSI_1-4 (&lt;100 m*)</u>
Count:	<b>11,909</b>	<b>11,409</b>	<b>3,486</b>	<b>3,396</b>
Minimum:	0	0	1	1
Maximum:	<b>857.829834</b>	<b>99.985001</b>	<b>718.936035</b>	<b>99.985001</b>
Sum:	405020.05598	232655.718402	100817.241362	68557.512587
Mean:	<b>34.009577</b>	<b>20.392297</b>	<b>28.920609</b>	<b>20.187725</b>
Stand Dev:	76.834624	11.022413	62.400914	10.166674
Median:	<b>18.867962</b>	<b>18.439089</b>	<b>18.788294</b>	<b>18.439089</b>

\* to exclude points on roads not represented on street layer (thus >100 m). Frequency Distribution →

\*\* 3,486 *confirmed* dengue cases – those with “CLASSI\_FIN” values of 1 (n=3,466), 2 (n=15), 3 (n=5), and 4 (n=0)



• **FINAL VALUE FOR ADDRESS LOCATOR:**

median of <100 m : **18.44 meters** = **60.50 feet**  
 (falls between standard offsets):  
**50 feet = 15.24 m**  
**100 feet = 30.48 m**



### III. ADDRESS LOCATOR

#### 1. Copy Brazilian Geocode and Locator files to system

- in Windows Vista, I went to **Program Files (x86) > ArcGIS** and copied folder **Geocode** and **Locators**, and added “**\_ORIGINAL**” suffix to both copies.
- I then copied the files from **02\_from ROBOTRON\_2010-03-28 > C\_ProgramFiles\_ArcGIS > Geocode** and **Locator** folders into the Program Files (x86) > ArcGIS folders of the same name, and skipped replacing all the files already present (effectively retaining original, US files and adding only the Brazilian ones).
- reformatted human data: “Dengue\_1998-2010\_11\_reformatted for geocoding.xlsx” (see HUMANS notes for details).

#### 2. Create CEP Address Locator

- **ArcCatalog > right-click destination folder > New > Address Locator > Choose an Address Locator Style: Endereço BR com CEP (File) > OK**

Name: **01\_Endereco BR com CEP File-based**  
 Description: [keep default: **Estilo brasileiro com CEP (File-based)**]  
 Reference data: ...DATA\STREETS\ADDRESS\_LOCATOR\03\_2010\_LOCATOR\Arruamento\streets\_P\_SAD69\_UTM23S.shp  
 Store relative path names [check box]

<u>Fields</u>	Tipo do logradouro:	<b>TIPO</b>
	Nome do logradouro:	<b>NOME</b>
	Numeração inicial esq:	<b>NUM_IE</b>
	Numeração final esq:	<b>NUM_FE</b>
	Numeração inicial dir.:	<b>NUM_ID</b>
	Numeração final dir.:	<b>NUM_FD</b>
	CEP esq.:	<b>CEP_LE</b>
	CEP dir.:	<b>CEP_LD</b>

<u>Input Address Fields</u>	<u>The field containing:</u>	<u>is recognized if it is named:</u>
[delete defaults, add following:] Street Zone		<b>ADDRESS_concatenated NU_CEP</b>

#### Matching Options

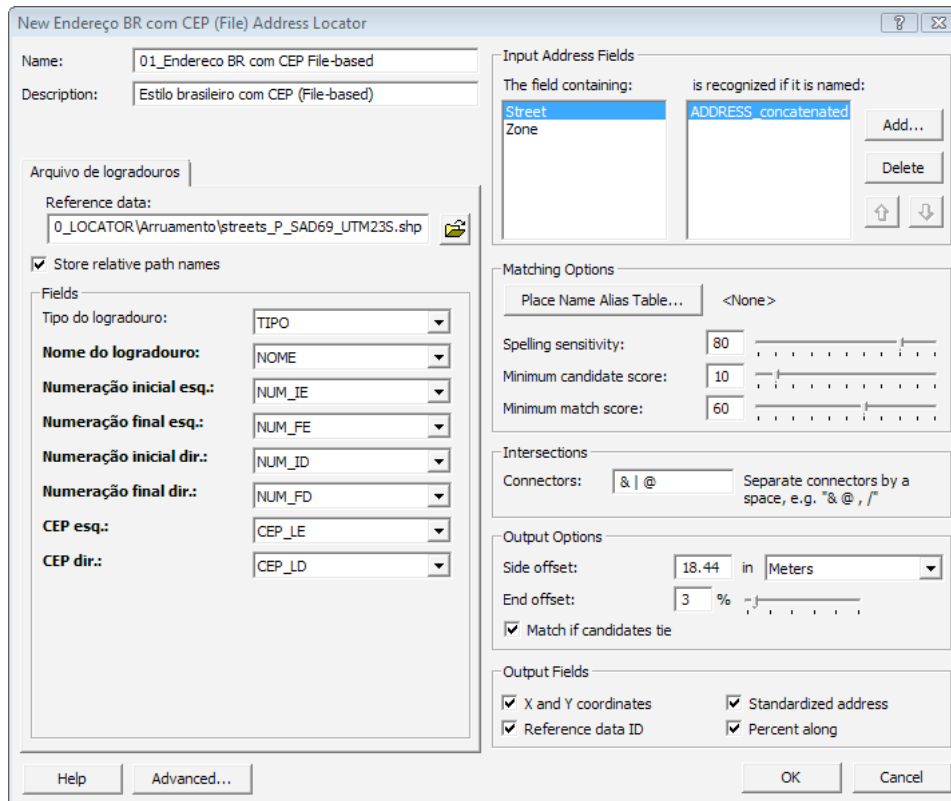
Spelling sensitivity:	<b>80</b>	[default]
Minimum candidate score:	<b>10</b>	[default]
Minimum match score:	<b>60</b>	[default]

Output Options

Side offset: **18.44 in Meters** [result from offset analysis]  
 End offset: **3** [default]  
 Match if candidates tie: [keep box checked]

Output Fields:

[check all four boxes]



**3. Create BAIRRO Address Locator:**

- **ArcCatalog** > right-click destination folder > New > **Address Locator** > Choose an Address Locator Style: **Endereço BR com CEP (File)** > OK

Name: **02\_Endereco BR com BAIRRO File-based**  
 Description: [keep default: **Estilo brasileiro com CEP (File-based)**]  
 Reference data:

...DATA\STREETS\ADDRESS\_LOCATOR\03\_2010\_LOCATOR\Arruamento\streets\_P\_SAD69\_UTM23S.shp

Store relative path names [check box]

Fields

Tipo do logradouro: **TIPO**  
 Nome do logradouro: **NOME**  
 Numeração inicial eq: **NUM\_IE**

Numeração final esq:	<b>NUM_FE</b>
Numeração inicial dir.:	<b>NUM_ID</b>
Numeração final dir.:	<b>NUM_FD</b>
CEP esq.:	<b>BAIRRO_LE</b>
CEP dir.:	<b>BAIRRO_LD</b>

<u>Input Address Fields</u>	<u>The field containing:</u>	<u>is recognized if it is named:</u>
	Street	<b>ADDRESS_concatenated</b>
	Zone	<b>BAIRRO</b>

Matching Options

Spelling sensitivity:	<b>80</b>	[default]
Minimum candidate score:	<b>10</b>	[default]
Minimum match score:	<b>60</b>	[default]

Output Options

Side offset: analysis]	<b>18.44 in Meters</b>	[result from offset
End offset:	<b>3</b>	[default]
Match if candidates tie:	<b>[keep box checked]</b>	

Output Fields: [check all four boxes]

4. **Create COMPOSITE Address Locator**

- **ArcCatalog** > right-click destination folder > New > **Address Locator** > Choose an Address Locator Style: **Composite** > OK

Name:	<b>03_COMPOSITE_01_and_02</b>
Participating Address Locators:	[click Browse button to right and select: <b>01_Endereco BR com CEP File-based</b> and then <b>02_Endereco BR com BAIRRO File-based]</b>

Input Address Fields:	<u>The field containing:</u>	<u>is recognized if it is named:</u>
	<b>Endereco</b>	<b>ADDRESS_concatenated</b>
	<b>CEP</b>	<b>NU_CEP</b>
	<b>BAIRRO</b>	<b>BAIRRO</b>

Click to select **01\_Endereco BR com CEP File-based** in Participating Address Locators:

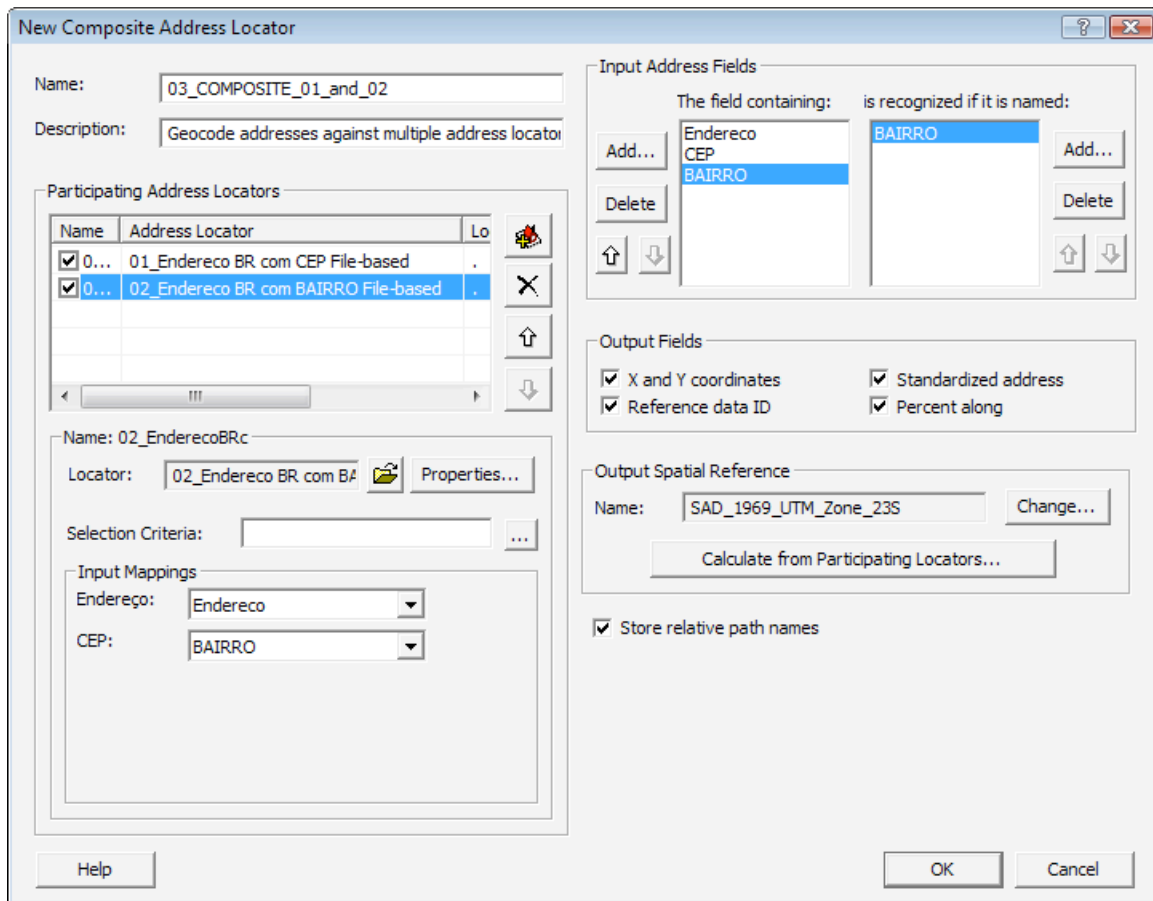
Input mappings:	Endereco:	<b>Endereco</b>
	CEP:	<b>NU_CEP</b>

Click to select **02\_Endereco BR com BAIRRO File-based** in Participating Address Locators:

Input mappings:	Endereco:	<b>Endereco</b>
	CEP:	<b>BAIRRO</b>

Output fields: [check all four boxes]

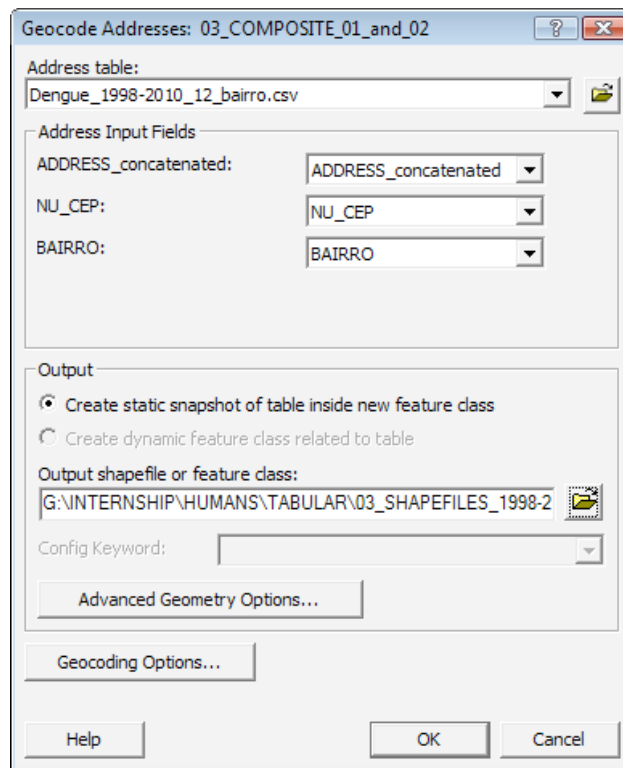
- Output Spatial Reference: [click **Calculate from Participating Locators**  
> select either]
- Store relative path names: [check box]



### 3. Geocode cases

- ArcMap > Add Data > **Dengue\_1998-2010\_12\_bairro.csv**
- Table of Contents' Source tab > Right-click .csv file > Geocode Address > Add > **03\_COMPOSITE\_01\_and\_02** > OK
- Address Input Fields: ADDRESS\_concatenated:  
**ADDRESS\_concatenated**  
NU\_CEP: **NU\_CEP**  
BAIRRO: **BAIRRO**
- Output shapefile or feature class:

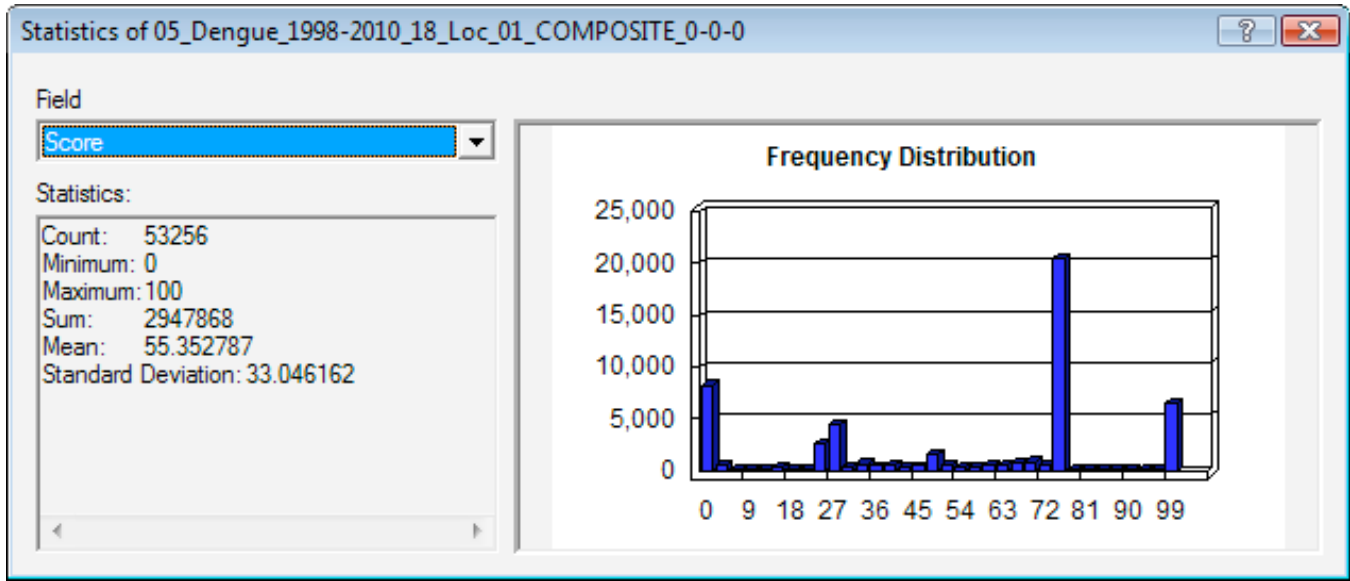
**01\_Dengue\_1998-2010\_12\_Loc\_03\_COMPOSITE\_80-10-60.shp**



# / Locator	Data	Parameters*	Matched	Tied	Unmatched	
01. 01_Endereco com CEP	18	80-10-60	30,006 (56%)	589 (1%)	22,661 (43%)	57%
02. 02_Endereco com BAIRRO	18	80-10-60	7,488 (14%)	124 (0%)	45,644 (86%)	14%
03. 03_Composite (01 + 02)	18	80-10-60	30,503 (57%)	598 (1%)	22,155 (42%)	58%
06. 03_Composite (01 + 02)	18	40-10-60	30,958 (58%)	606 (1%)	21,692 (41%)	59%
04. 03_Composite (01 + 02)	18	40-10-30	35,866 (67%)	777 (1%)	16,613 (31%)	68%
15. 03_Composite (01 + 02)	18	80-10-23	35,732 (62%)	903 (2%)	16,621 (31%)	64%
14. 03_Composite (01 + 02)	18	60-10-23*	37,544 (70%)	940 (2%)	14,772 (28%)	72%
16. 03_Composite (01 + 02)	18	50-10-23	38,485 (72%)	1,071 (2%)	13,700 (26%)	74%
10. 03_Composite (01 + 02)	18	40-10-23*	39,114 (73%)	1,558 (3%)	12,584 (24%)	76%
13. 03_Composite (01 + 02)	18	20-10-23	39,474 (74%)	3,529 (7%)	10,253 (19%)	81%
07. 03_Composite (01 + 02)	18	40-10-15	39,420 (74%)	1,573 (3%)	12,263 (23%)	77%
08. 03_Composite (01 + 02)	18	20-10-15	39,800 (75%)	3,551 (7%)	9,905 (19%)	82%
09. 03_Composite (01 + 02)	18	20-5-15	39,800 (75%)	3,551 (7%)	9,905 (19%)	82%
11. 03_Composite (01 + 02)	18	20-5-5	40,073 (75%)	3,568 (7%)	9,615 (18%)	82%
12. 03_Composite (01 + 02)	18	5-5-5	39,893 (75%)	4,618 (9%)	8,745 (16%)	84%
05. 03_Composite (01 + 02)	18	0-0-0	40,352 (76%)	6,490 (12%)	6,414 (12%)	88%

\* Parameters: Spelling sensitivity – Minimum candidate score – Minimum match score

2. **MMS Distribution analysis:** for 0-0-0, distribution of scores (compare vis-à-vis Minimum Match Score). Used 0-0-0 because for Unmatched, all records < Minimum Match Score are set to zero, apparently)



\* second group of peaks is at Score = **24** (n=2,519), another at **28**. So, set MMS at 23.

- If I remember correctly, the minimum match score for the WNV dead bird data (circa 2005, 2006) was also **30**

**3. Analysis of Tied's:**

- 20-10-23: the "Matches" around 28 and lower were not matches! don't use!
- 40-10-23: were better, but still off.
- **60-10-23: all were matches, very good!**
- 50-10-23: mostly matches, a few incorrect however.

**4. Manual rematching**

- Address numbers within 100 were cool. All of potential matches were within 100 though.
- Rematched half (n=7,500) of Unmatched (n=14,772). Matched **252** (1% of total).
- Found that all candidate addresses with score of 10 or higher (may not have actually been any <10) were Matched, therefore Step 5:

**5. Automatic rematching**

- Geocoding Options button (bottom left) > set both address locators in COMPOSITE to **60-10-10**

- This matched an additional **252** (246 Matched, 6 Tied), bringing the final total to:

<u>Matched</u>	<u>Tied</u>	<u>Unmatched</u>	<u>M+T</u>
38,042 ( <b>71.4%</b> )	946 (1.8%)	14,268 (26.8%)	38,988 ( <b>73.2%</b> )

- recommend using **60-10-10** in the future

- **from 19**

total geocoded:	<b>38,988</b>	<b>06a_FINAL_HUMANS..._n_38988.xls</b>
total geocoded 1-4:	<b>13,546</b>	<b>06b_FINAL_HUMANS..._n_13546.xls</b>





Right: 224236.168202903  
Bottom: 7639495.44069285

height:  $7666274.3256114 - 7639495.44069285 = 26778.88491855$   
width:  $224236.168202903 - 197457.283284349 = 26778.884918554$

### 3. Convert to raster

- **Spatial Analyst toolbar > Convert > Features to Raster**

Input features: DYCAST\_RP\_polygon  
Field: Id  
Output cell size: 100  
Output raster: **DYCAST\_RP\_ras**

### 4. Sample raster (add X,Y coordinates)

- **ArcToolbox > Spatial Analyst Tools > Extraction > Sample**

Input rasters: DYCAST\_RP\_ras  
Input location raster or point features: DYCAST\_RP\_ras  
Output table: **DYCAST\_RP\_centroids**  
Resampling technique (optional): [keep as NEAREST]

### 5. Display centroids

- Add Data > **DYCAST\_RP\_centroids**
- Table of Contents > Source tab > right-click new table > **Display XY Data**

X Field: X  
Y Field: Y  
Coordinate System of Input Coordinates > Edit > Select > **South American 1969 UTM Zone 23S.prj**

- Display tab > right-click new shapefile > Open Attribute Table > right-click “DYCAST\_RP\_RAS” > **Delete Field**. Repeat for column, “DYCAST\_RP\_RAS\_1” and close out window.
- right-click shapefile > Data > Export Data > **effects\_poly\_centers\_projected.shp**

- click Yes to add exported data to the map as a layer.

## 6. Convert centroids to raster

- ArcToolbox > Conversion Tools > To Raster > **Point to Raster**

Input Features: effects\_poly\_centers\_projected  
Value field: FID  
Output Raster Dataset: **effects\_rast**  
Cell assignment type: [keep default, MOST FREQUENT]  
Priority field (optional): [keep default, NONE]  
Cellsize: 100

## 7. Convert raster to shapefile

- Spatial Analyst toolbar > Convert > **Raster to Features**

Input raster: effects\_ras  
Field: VALUE  
Output geometry type: Polygon  
Generalize lines: [keep box checked]  
Output features: **DYCAST\_RP\_cells\_n\_71824.shp**

## 8. Add grid to PostgreSQL

- **pgAdmin III**: in Object Browser, navigate to: Servers > PostgreSQL 8.4 > Databases > dycast > Schemas > public > **Tables**
- right-click existing **effects\_poly\_centers\_projected** > **Delete/Drop**
- EPSG code for **SAD69 UTM Zone 23S = 29193**  
<http://spatialreference.org/>

- **Plugins** menu > **PostGIS Shapefile and DBF loader**:

Shape File: **effects\_poly\_centers\_projected**  
SRID: **29193**  
Destination table: **effects\_poly\_centers\_projected**  
Geometry Column: [leave as “the\_geom”] > Import

- Close window after import. Right-click Tables > **Refresh**

- Expand **effects\_poly\_centers\_projected** > **Columns**:
  - right-click “**gid**” > Properties > Name: change to “**tile\_id**”
  - right-click “**x**” > Delete/Drop > Yes
  - right-click “**y**” > Delete/Drop > Yes
- close pgAdmin III.

---

## Appendix IV. Protocol – DYCAST, ArcMap analysis

---

### I. ADD DATA

---

#### 1. Convert to .tsv

- **Excel:** open DBF from geocoded humans shapefile. Reformat columns as follows (add “id” and “species” columns):

id	onset_date	longitude	latitude	classi_fin
1	1/4/1998	207049.01180548100	7658232.68130018000	1

- Saves As > **Text (Tab delimited)**. It will save as a .txt file.
- In DYCAST > **inbox** folder, duplicate **dycast\_export\_2008-04-04.tsv** file. Rename as **dycast\_export\_2010-04-05.tsv** and open up in WordPad or Notepad (Windows). Select all and delete everything.
- Open up new .txt file in Notepad or Wordpad. Copy-paste everything into the .tsv file. Save file. [DYCAST can’t read anything but .tsv files]

#### 2. Run DYCAST procedure.

- #### 3. Note: to add new humans (“birds”), open pgAdminIII > Tables > **dead\_birds** > View Data > right-click row headers and **Delete**.

This will automatically delete the same records from the **dead\_birds\_projected** table.

Close the program and run DYCAST, selecting and loading birds as per usual.

---


## II. ANIMATIONS

---


### I. CREATE PERSONAL GEODATABASE [ArcCatalog]

1. within the folder viewer at the left, create new folder, **DYCAST\_RISK**
2. right-click folder > New > **Personal Geodatabase**
3. Rename geodatabase based on risk model parameters:  
“**600-100-4-28\_TH05.mdb**”
4. right-click geodatabase > Import > **Feature Class (single)**  
  
Input Features: **dycast\_rp\_cells\_n\_71824.shp**  
Output Feature Class: **DYCAST\_GRID**
5. right-click geodatabase > Import > **Table (multiple)**
6. For “Input Table,” navigate to and select all the risk files:  
  
**riskYYYY-MM-DD.dbf** [may take awhile to load]  
  
click Add > OK
7. Open geodatabase. Right-click > Copy/Paste the first (earliest) risk file. This will be the target table in the following steps; rename this as:  
  
**riskYYYY\_MM\_DD\_to\_YYYY\_MM\_DD**[cannot use dashes in name]
8. ArcToolbox > Data Management Tools > General > **Append**:  
  
Input Datasets: select all the *other* tables from the new geodatabase (not the original dbf files, which will be displayed initially in the browser)  
Target Dataset: the target table from the previous step.  
Schema Type: TEST
9. Export table for use in accuracy calculation (RESULTS Notes)
  - right-click target table > Export > **To dBase (single)**:  
  
Output Location: **DYCAST\_RISK** folder  
Output Table: **600-100-4-28\_TH10\_risk2005-07-01\_to\_2007-04-09.dbf**

## II. ADD GEODATABASE TO TRACKING ANALYST [ArcMap]

1. Click the “Add Temporal Data” button: 
2. select “A feature class and a separate table containing temporal data that this wizard will join to the feature class” (1)
3. for “Choose the input feature class” (2), navigate to geodatabase and select **DYCAST\_GRID**
4. for “Choose the input table” (3), navigate to geodatabase and select the table with the appended years > **risk2005\_07\_01\_to\_2007\_04\_18**
5. for “Choose the field containing the data/time from:” (4), select:  
Or the input table: **DATE\_** [default]
6. Click “Next”
7. for “Choose the fields to base the join on:” (5), choose:  
Join Field in Input Feature Class: **GRIDCODE**  
Join Field in Input Table: **ID**
8. Click “Finish”
9. repeat steps 1-8 for every other years/thresholds

## III. ADD HUMANS IN TRACKING ANALYST

1. Click the “Add Temporal Data” button: 
2. select “A feature class or shapefile containing temporal data.” (1)
3. for “Choose the input feature class” (2), navigate to and select the geocoded case shapefile
4. for “Choose the field containing the data/time from:” (3), select “**onset\_date**”
5. Leave (4) as default, “<None>”
6. Click “Finish”

#### **IV. SET PREFERENCES**

[ArcMap]

1. Right-click temporal layer(s) and select: Properties > Symbology tab
2. In the “Show:” window at the top-left, check the “**Time Window**” box.
3. In the new “Time Window” box, type **28** for **Period** and select **Days** for **Units**  
> OK [for HUMANS; for risk do 1 day]

---

### III. ANALYSIS

---

#### I. JOIN

- **Create joined shapefile of humans with cells (one-to-one)**

ArcToolbox > Analysis Tools > Overlay > Spatial Join:

Target Features: 08\_FINAL\_HUMANS...n\_13546\_...ALL\_FIXED  
Join Features: DYCAST\_RP\_CELLS\_n\_71824  
Output Feature Class: 08b\_FINAL...WITH\_CELLS.shp  
[no dashes]  
Join Operation: **JOIN\_ONE\_TO\_ONE**  
Match Option: IS\_WITHIN

#### II. DATA

- ArcCatalog > export to DBF, then open in/convert to Excel .xls.
- sort **RISK\_TH10\_7-01-05\_to\_6-30-2006** by **ID** then **DATE\_**, both Ascending.
- copy/paste entire **ID** column to a new worksheet, then Data > Subtotals > **Count** (this will take awhile). This gives the number of days lit per cell, as well as a list of unique ID numbers for the subsequent lookup table.
- click on the “2” box at the very top-left to collapse the count records. Then copy/paste both columns into a blank Word document, then copy/paste these results back into Excel, in a new worksheet.
- Select column A and **Find/Replace All** “[space]Count” with “[blank]” to turn ID values back to numbers. Add “**ID**” as column A heading, and change column B heading to “**DAYS\_LIT**”
- Save file, then Save As “...\_STATS\_FINAL”
- Insert a new column between ID and DAYS\_LIT. Name it **FIRST\_LIT** and fill the values with the formula:

**=VLOOKUP(A2,[ID and DATE columns from original worksheet],2,FALSE)**

which should read something like this:



**=VLOOKUP(A2,'600-100-4-28\_TH10\_risk2005-07-0'!\$A\$2:\$B\$461596,2,FALSE)**

*\*\*\* remember to put the dollar signs in to lock the array!*

This will grab the first date listed from the original page (which, since they're sorted by ID then DATE, will be the earliest date lit). Change the cell format from number to date.

- copy-paste these three columns into a new workbook (i.e., “**RISK\_TH10**” in the composite Excel file, “**ACCURACY\_STATS\_**”, as **Values Only**).

### III. STATISTICS

- Run descriptive statistics using **StatPlus** stand-alone program (or Analysis Pak if running PC version of Excel) to calculate count, mean, std dev, max, and mode of **DAYS\_LIT**. (watch out for check box denoting whether you're including column title). Copy/paste results into Excel document next to raw data, then into compiled results worksheet, “**STATS\_FINAL**”
- Make sure “**DATA\_08b\_H\_w\_CELLS**” worksheet is sorted by **GRIDCODE** then **onset\_date**, both ascending.
- Create new B column in , “**HUMANS\_first onset**”. Fill the values with:

**=VLOOKUP(A2,DATA\_08b\_H\_w\_CELLS!\$A\$2:\$B\$4691,2,FALSE)**

*...again, make sure to enter \$ signs, and also that onset\_date is column 2*

- Change cell format to Date, then copy/paste result into a new column as **Values Only**. Delete original column with formulas.
- Sort column descending, and count number of “#N/A”s. This is the value for the **FALSE POSITIVES** (b cell) in the 2x2 confusion matrix in **STATS\_FINAL** worksheet.
- Sort **RISK\_TH10** data again but by **ID**, ascending. In **DATA\_08b\_H\_w\_CELLS** worksheet, add new B column, “**TH10\_first lit**” and **VLOOKUP** values from respective (“**RISK\_TH10**”) worksheet. Then duplicate column into new one as **Values Only**, then delete original.
- Sort by **GRIDCODE** then **onset\_date**.

- Create DUPLICATE column K:

**=IF(F2=F3,IF(B2=B3,0,"DUPLICATE\_w\_diff\_ONSET"),)**

to flag duplicate records.

- Create new C column, "**TH10\_days PRED**". Use formula:

**=IF([DUPLICATE]2=0,[onset\_date]2-[TH10\_first lit]2,"DUPLICATE")**

such as:

**=IF(K2=0,F2-D2,"DUPLICATE")**

this will generate the # of days predicted for the first cases within a cell.

- Change the column cells' format to Number with no decimal places, then copy/paste into new column as Values Only, delete original. Manually verify accuracy of values.
- Sort days PRED column Ascending. Color-code negative numbers (misses=false negatives) as red, zeroes and positive numbers (true positives) as green, ignore DUPLICATE's (but count them to verify total), and #N/A's as red (never lit=false negative)

Enter counts below column, transfer values to STATS\_FINAL worksheet confusion matrix.

- Regarding the number of total cells: select dycast cells that intersect street reference data with a 618.44 meter buffer (spatial domain + geocoding side offset. **(N = 25,487 cells)**)
- TRUE NEGATIVES (d cell) in confusion matrix should be calculated as the remainder needed to bring total to 25,487 cells.
- Run descriptive statistics (StatPlus) on **\_days PREDICTED** column for those predicted values (cells with 0 or positive values). Plot TH10 and TH5 on ROC graph.