ArboMAP: Arbovirus Modeling and Prediction to Forecast Mosquito-Borne Disease Outbreaks

User Guide (v3.1)

Justin K. Davis and Michael C. Wimberly
(mcwimberly@ou.edu)
Geography and Environmental Sustainability, University of Oklahoma

Updated June 16, 2021

Contents

Introduction

What you need to know about ArboMAP

ArboMAP is a program used to model and predict cases of vector-borne disease. It has been developed and tested with human cases of the mosquito-born West Nile virus. However, ArboMAP can be adapted to work with any data sets that meet the following conditions:

- There are multiple years of infection data from years in which the pathogen can be considered endemic. Introductory years in which the pathogen is probably rapidly exploiting naive populations are likely not representative of transmission dynamics in subsequent years, and should be removed from the dataset.
- The pathogen has distinct transmission and quiescent seasons, with an initial growth phase at the beginning of the transmission season when the pathogen is beginning to spread after a period with minimal to no transmission. In the current version, ArboMAP assumes that the transmission season occurs during a single calendar year and ceases during the boreal winter. In settings where a transmission season crosses the December-January boundary, the code will require modification.

- There is reason to believe that incidence responds to measurable environmental indices, typically including temperature and some measure of moisture in the environment (e.g. precipitation or humidity)
- Cases of disease are assigned to districts (e.g., counties) and cases are not too rare - the best situation is where every modeled district has had at least one case over the period of study.
- Some measure of pathogen in the environment is available; e.g. here we use the rate at which pools of mosquitoes test positive for the virus.

ArboMAP is designed to facilitate statewide forecasting of West Nile virus risk at the county level, and this guide is primarily focused on this implementation. However, with some modification, the code could be used with different geographic units (for example, zip codes or states) and at different spatial extents (for example, an individual county of the entire United States).

The ArboMAP user guide consists of four main sections. The first section describes how to install the necessary software needed to run the system. The second section describes the steps that need to be taken every year before the beginning of the WNV season to prepare for forecasting. The third section description the steps that need to be taken each time a new forecast is generated. The final section describes the outputs produces by ArboMAP and explains how to interpret the forecast charts and maps.

What you need to know about WNV and modeling

West Nile virus (WNV) in South Dakota is our example in this document. WNV circulates primarily among mosquitoes and birds, but occasionally spills over into human or other hosts. In humans, the majority (~80%) of infections show no symptoms and are not diagnosed except perhaps accidentally during blood donation. Some small proportion of individuals will develop symptoms such as headache, fever, and rash, but in about one out of every 100 infected individuals, the disease becomes neuroinvasive and can cause debilitating illness and death.

WNV was first introduced to the U.S. in 1999 on the east coast, after which it spread in a westward-moving wave across the country The early years produced thousands of cases as WNV invaded new regions. Birds were immunologically naive, so the virus decimated entire avian populations and spilled over into humans in an outbreak that lasted several years in some states.

In South Dakota, there were more than a thousand reported cases in 2002-2003 (Wimberly et al. 2012), which implies that many more were actually infected, since most infected individuals do not show symptoms and are never diagnosed. After a number of years of active transmission, case numbers declined and it looked like the disease might vanish from circulation after 2011, in which there were only two cases reported in South Dakota.

However, in 2012 there was another outbreak, with hundreds of cases in South Dakota (Wimberly et al. 2013). Because of these fluctuations in annual case counts, there is a need to predict the magnitude and locations of WNV in advance to support proactive mosquito control and disease prevention activities. Thus, a key scientific question is whether there are there any environmental or entomological indicators that could tell us, in advance, how much WNV transmission is likely to occur in the current year.

Multiple studies conducted in South Dakota have confirmed that vector mosquitoes and human WNV cases are sensitive to fluctuations in environmental variables, particularly temperature (Chuang et al. 2011, 2012a, 2012b; Hess et al. 2018; Wimberly et al. 2008, 2014). Our research has shown that relatively simple statistical models, relying on weather and mosquito infection data, can be used to predict the risk of infection on a district-week basis (Davis et al. 2017, 2018). For more background, you will want to consult relevant journal articles, which are listed at the end of this document.

Setting up for the initial run

A note on prepackaged data

ArboMAP comes packaged with example human, mosquito (including spatial strata), and weather data. The weather data are real, but the human and mosquito data were generated randomly according to model estimates from our WNV study in South Dakota. Hence, the data appear similar to real human and mosquito data, but are synthetic and *should not be used* for any actual scientific study. They are included merely to make sure that all software is installed and settings are correct; you will probably want to run the system first with these synthetic data before trying your own data.

Some data you'll need

ArboMAP uses data on mosquito infection rates as predictors in the model. We assume that such data are acquired through mosquito surveillance programs implemented at the state, county, and or municipal level, and that users of ArboMAP will have access to these data. For example, during the initial implementation of ArboMAP in SD, we accessed the data through a web-based data portal to which which participants around the state had uploaded their mosquito infection data. Information is required about the county where each mosquito pool was collected, the date that it was collected, and whether it tested positive or negative for WNV.

You will need weather data so that relationships between disease and environment can be modeled. To forecast WNV in the United States, ArboMAP uses the gridMET dataset, which provide daily gridded data on meterological variables generated through fusion of the NLDAS and PRISM datasets. We provided a web-based application that can be used to download these data for free from Google Earth Engine (GEE). GEE is a cloud-based platform for processing satellite remote sensing images and other geospatial datasets. The application allows users to select a state and time range, automatically processes the meterological data, and provides daily, county-level summaries in a comma-delimited text format that can be imported into ArboMAP.

To access GEE, you will first need a gmail account at **mail.google.com**, so that you can access Google Drive documents at **drive.google.com**. Next, visit GEE at **earthengine.google.com** and click on "Sign Up" to request an account. You receive confirmation at your gmail address, and any downloads of weather data will be stored in the Google Drive of the account.

Finally, you will need outcome data, usually human cases. As with the mosquito data, we assume that the users of ArboMAP will have access to these data. ArboMAP models the current year based on all previous years, and does not rely on case counts in the current year because these cases are typically reported with time lags of weeks to months. The county of residence (or transmission) and a date (typically the onset of symptoms) are required for each case.

Some software you'll need

R

R (**www.r-project.org**) is a statistical programming language that runs all of our analyses and produces reports and documentation, including this document. It is free and has a wide variety of packages built by users all around the world to do essentially any statistical analysis you can imagine.

It is easiest to download R from CRAN (**cloud.r-project.org**). Click on the link for "Download R for Windows." Choose "base," then "Download R for Windows." Run this file and install R on your system. Use the default settings for the installer.

RStudio

RStudio (**www.rstudio.com**) is a user-friendly GUI for the statistical programming language R that greatly simplifies a number of tasks for the programmer. Navigate to the site, click on Download, and choose the RStudio Desktop (Open Source License) - that is, the free version. Run the appropriate installer, very likely the Windows Vista/7/8/10 installer.

Run RStudio and close it at least once after you've installed MiKTeX as described in the next section. RStudio may not be able to find MiKTeX unless it's had a chance to search.

MiKTeX

MiKTeX (**www.miktex.org**) is an implementation of LaTeX, which is what allows us to automatically produce PDF reports at the end of all the calculations. Navigate to the site, click on Download -> Windows and then download and run the installer. Default options will suffice, but make one change: "Install missing packages on-the-fly" should be changed to "yes." You will initially install an incomplete copy of MiKTeX, and it will be updated automatically whenever you run the code. This means that the first time you run the code, it might take quite a while.

An alternative is to choose the Net Installer from the "All Downloads" page, and to choose "Complete Installation." This will take a great deal of time, and you should only do it on a fast connection, but it means you will only ever have to do this once. Expect that this will take a few hours - if you can afford the time to do this, it's worth getting all the downloads done. You will probably be asked to select a source/server/mirror from which to download your files - you can choose the 0-Cloud, which will automatically find a fast download site for you, or you can manually pick somewhere close to you physically.

Setting up the directory structure

ArboMAP requires a variety of files to run, and the following directory structure is the easiest way to organize everything that's necessary. If you download the repository and run the main ArboMAP code in RStudio using ArboMAP.Rrpoj, you will draw from data sources already in place and will not need to modify the contents of these directories. By default, ArboMAP will expect the following folders to be subfolders of the same home directory of your RStudio project.

- **human case data**: Contains the file with human case data
- **mosquito data**: Contains the file with mosquito data
- **strata**: Contains the file with spatial strata for the mosquito data
- **weather data**: Contains the files with weather data
- **pictures for setup document**: Contains images necessary to compute the user guide, but not required to run the main ArboMAP

Files and settings

The ArboMAP-Main-Code.Rmd file, which you will open in RStudio, contains a preface (shown below) with settings that tells the program where to find the various pieces of data used to make predictions. Errors in these filenames are a common reason that the code will refuse to run.

- `state_name` is the name of the state being modeled. It is included in the title of the report generated by ArboMAP.
- `state_code` is the two-letter code of the state being modeled.

- `forecast_date` tells the program which week you'd like to obtain predictions for. Weeks in this program begin on Sunday and run until the next Saturday, and the `forecast_date` date will automatically be rounded to the previous Sunday. For example, if 2018-07-16 happens to be a Monday, then the week examined is 2018-07-15 (Sunday) through 2018-07-21 (Saturday).
- `human_file` is the name of the file that contains human case data.
- `mosquito_file` contains the mosquito testing data.
- `stratification_file` specifies the allocation of counties to strata for the spatially stratified version of the mosquito infection rate model.
- `predictor_v1` and `predictor_v2` are the names of the two meteorological predictor variables.
- `mosquito_model` is a code for the type of mosquito infection rate model to be used: MIGR = Mosquito infection growth rate, stratifiedMIGR = spatially stratified version of MIGR, MII = Mosquito infection intercept (constant term from the mosquito infection growth rate equation, stratifiedMII = spatially stratified version of MII, AUC = area under the mosquito infection growth curve, simpleratio = percent positive mosquito pools).
- `min_human_year` is the earliest year of available human case data.
- `max_human_year` is the most recent year of available human case data (usually the year before the current forecast year).
- `min_mosquito_year` is the earliest year of available mosquito infection data.
- `max_mosquito_year` is the most recent year of available mosquito infection data (should be the current year during which the the forecast is being made).
- `min_weather_year` is the earliest year of available weather data.
- `max_weather_year` is the most recent year of available weather data (should be the current year during which the forecast is being made).
- `min_desired_year` is the earliest year to include when fitting the model (typically the first year when human data and either mosquito or weather data are available).
- `max_desired_year` is the most recent year to include when fitting the model (typically the current year when generating forecasts).
- `case_trim_alpha` tells ArboMAP to discard the earliest/latest alpha% of the human cases in all years. This prevents bad data (e.g. infections reported in December but actually contracted in July) from causing numerical difficulties in the analysis.
- `compyear1` and `compyear2` indicate which two years in the past to compare the current year to in some of the charts generated by ArboMAP.
- `resample_mosquito` indicates whether a resampling procedure should be used to construct realistic seasonal time series based on seasonal totals.
- `resample_file` must be provided if the previous option is `TRUE`. Otherwise, this input is ignored.
- `data_path` is the path to the main folder that contains the data subfolders. But default the value is set to ".", which references the working directory of the current R project. However, this value can be changed if the user wants to work with different datasets stored in different locations.

Choosing your models

ArboMAP predicts WNV cases using "big additive models" run with the `bam()` function from the mgcv library. Predictor variables include weather variables with effects modeled as distributed lags, and mosquito infection rates modeled using one of several techniques. There are a variety of other modeling options:

- Models can be based on 1) untransformed weather data, or 2) weather anomalies, calculated as the difference between each daily value and its long-term daily expectation, combined with a cyclical seasonal trend.

- Distributed lags can be 1) a single, fixed set of distributed lags, or 2) seasonally varying so that the lag functions can change over the course of the WNV season.

- Smoothed functions can be modeled using 1) cubic splines, or 2) thin-plate splines.

Models to be used in an ArboMAP run are specified in a "models.txt" file that must be present in the working directly of the RStudio project. Each line in this file contains two comma-separated strings. The first contains a code name for the model and the second contains the specification of the R model formula. In the file provided with ArboMAP, the naming convention describes whether cubic (cub) or thin-plate (tp) splines are used. Then, the distributed lags are either fixed (fx) or seasonally-varying (sv). Then, we either use anomalized (anom) or non-anomalized (nonanom) environmental data.

In general, we expect that models based on thin-plate splines will be more stable and perform better than models based on cubic splines. However, this may not always be the case, and fitting models based on thin-plate splines typically takes longer. We have found that models based on weather anomalies tend to outperform those based on transformed weather variables, though again this is not necessarily always the case. Models with seasonally-varying coefficients can be more effective when environmental sensitivies are different early in the season (when virus amplification is occurring in bird populations) versus later in the season (when infected mosquitoes are biting humans). However, the seasonally varying models are more complex, take more time to fit, and may be more likely to overfit to the training dataset.

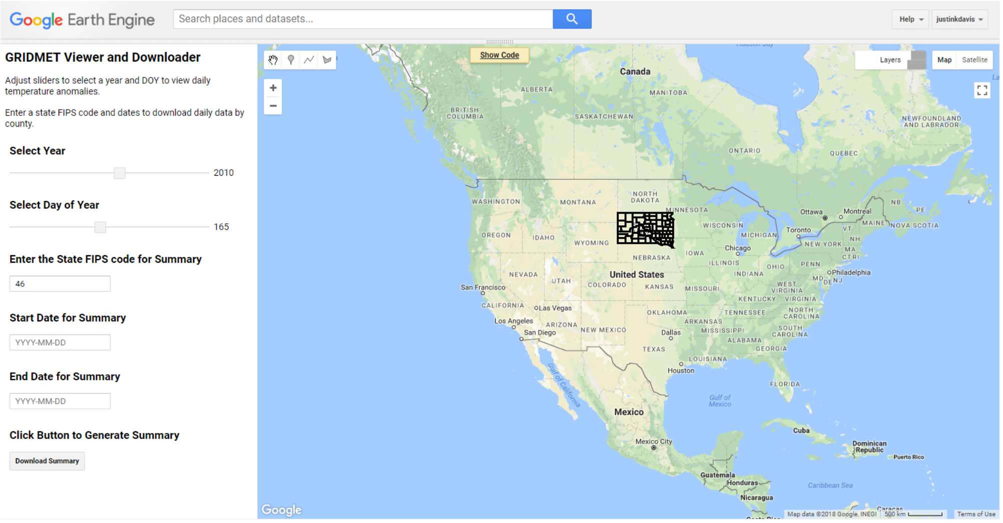
Other important modeling choices include determining which meteorological variables to use as predictors, and which type of mosquito index to use. Based on previous research, we have found that best results are obtained by combining temperature with a variable related to moisture (vapor pressure deficit, relative humidity, or precipitation). It is also necessary to choose a technique for modeling the mosquito infection data to generate entomological risk indices. In most situations, the mosquito infection growth rate (MIGR) is the best model of entomological risk, particularly in the early part of the WNV season. In situations where the mosquito infection rates does not exhibit unimodal growth in the early season, the mosquito infection intercept (MII) or the area under the mosquito infection growth curve (AUC) can be used as alternatives. For MIGR and MII, there is the also the option of implementing a spatially stratified version where multiple can be use in different parts of the state.

The report generated by ArboMAP provides fit statistics for each model formula included in the run. Only one set of meteorological variables and one mosquito infection model can be selected for each run. Therefore, multiple runs must be made to compare different optiosn for these settings. A suggested strategy is to start with multiple runs that include all the model formulas combined with environmental variables and mosquito infection models. Then, a smaller subset of the best-fitting models can be selected to use for routine forecasting.

Obtaining data at the beginning of the season

Obtaining the gridMET data

ArboMAP requires weather data to make its predictions. You will use likely Google Earth Engine (GEE) to download historical weather data beginning a full year before your human case data (e.g. if you begin modeling in 2001, you will need to download all data beginning from 2000). If you want to download all available historical data, it will not hurt the calculations; simply begin in 1950). You will only need to download historical data once; every other time you run the program, you will only need to download the most recent year's weather data from GEE. Navigate to https://earthengine.google.com/ and sign up for a GEE account if you do not have one already.



GRIDMET Viewer and Downloader

Adjust sliders to select a year and DOY to view daily temperature anomalies.

Enter a state FIPS code and dates to download daily data by county.

Select Year 2010

Select Day of Year 165

Enter the State FIPS code for Summary 46

Start Date for Summary YYYY-MM-DD

End Date for Summary YYYY-MM-DD

Click Button to Generate Summary

Download Summary

In the ArboMAP repository there is a .js file (e.g. GRIDMET_downloader_v2_1.js). This contains the latest version of the GEE Javascript script. Navigate to https://code.earthengine.google.com/ and in the “New Script” center section of the page, copy and paste the text of the script. Click on save, and name it the same as the text file (or another name of your choosing). Alternatively, use the “New . . . File” option in the left-hand Scripts pane to create an empty script. Click on the new script in the Scripts pane and then copy in the contents of the .js file. You may need to click on the textured bar above “Show Code” and drag down to view the code first.

Then click on “run” to render the GUI. You don’t need to modify any of the code, but this will allow you to see the Tasks tab, which will allow you to download the data. The start and end dates are automatically set to provide data for the most recent month. You can change these dates if you need to download data from a different time period.

You can change some of the other options, if you’d like to modify your download. If you’re not using the data for South Dakota, for example, which has FIPS code 46, you can choose another state from the list of codes (**click here for codes**). You will need to use the two-digit code (e.g. Colorado is 08, not 8).

Click on the button that says"Click to Generate Task for Download Summary." This will create a task under the Tasks tab. The name of the Task will be automatically based on the state and date range of the download. Click on Run. If you do see the Tasks or any of the code, click on the textured bar directly above the map and drag down. This should reveal the code window and the Tasks tab.

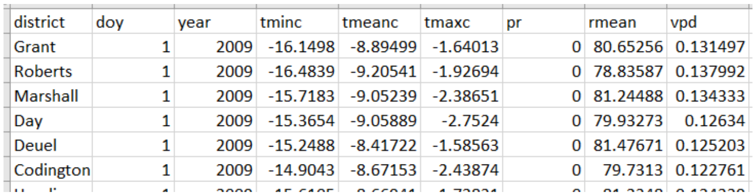
When setting up the data for a new state, the user must start by downloading historical data for model calibration - the weather data should extend at least one year before the earliest human case data. The application can download multiple years of data. When making large downloads, processing can take half an hour or more depending on who else is using the system, and very rarely your job can be cancelled - you might receive an error about "too many objects," in which case you should try again later or split the download into individual years. Once it’s done, there will be an new file in your Google account’s drive at **drive.google.com**.

Optional: You can access your drive online at https://www.google.com/drive/. Or you can download a desktop app https://www.google.com/drive/download/ and use it like a folder on your computer.

Download this file (right-click -> download) and save it in a directory with all the other weather data; by default this is R project subfolder "./weather data" ("/ArboMAP/weather data/"). When making forecasts, weather data are usually updated on a weekly basis. These updates are stored as extra CSV files and, along with the historical download, give you a complete picture of the daily weather in every district.

Note that it doesn’t matter how many times you download a date’s weather data, as long as it appears at least once in the data. The code will automatically take the most recent measurement for that date and additional measurements will not affect the predictions.

If you're not using GEE Although GEE is recommended for a variety of reasons, you might obtain your environmental data from elsewhere. Below is format expected of any CSV file that you might instead supply. District is a human-readable name identifying the area where the weather measurement was taken. The "doy" indicates the day of the year (1-366). The remaining columns are measurements of environmental conditions. If your measurements are not daily (e.g. 7-day precipitation totals), they will need to be resampled to that temporal resolution.



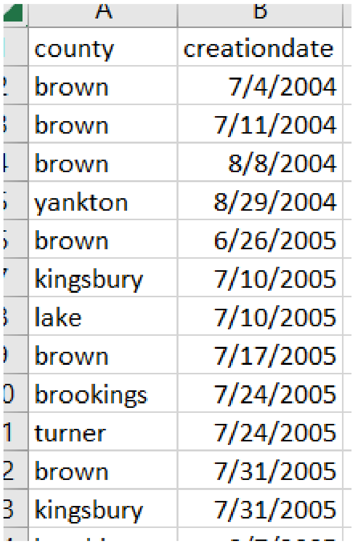
district	doy	year	tminc	tmeanc	tmaxc	pr	rmean	vpd
Grant	1	2009	-16.1498	-8.89499	-1.64013	0	80.65256	0.131497
Roberts	1	2009	-16.4839	-9.20541	-1.92694	0	78.83587	0.137992
Marshall	1	2009	-15.7183	-9.05239	-2.38651	0	81.24488	0.134333
Day	1	2009	-15.3654	-9.05889	-2.7524	0	79.93273	0.12634
Deuel	1	2009	-15.2488	-8.41722	-1.58563	0	81.47671	0.125203
Codington	1	2009	-14.9043	-8.67153	-2.43874	0	79.7313	0.122761

Obtaining the human case data

You will only need to update the human case data file once at the beginning of each season. For example, to make WNV forecasts for 2018, the model is calibrated using human case data through the end of 2017, and 2018 case data are not incorporated until data are finalized after the end of the 2018 WNV season. Only two pieces of information are needed for every human case: a date and a district. The date is the symptom onset date (in the case of clinical cases) or the date of blood donation (in the case of viremic blood donors). The district is the patient's district (e.g. county) of residence. Each row represents a single case.

The CSV file you use should that contains the human case data should have a unique column named "date" that is formatted MM/DD/YYYY, and a column named "district" that contains standard names for location of residence (or transmission, if available). The names of districts are simplified by the code, so that any of the following are equivalent: BROOKINGS, Brookings, brookings, Brookings County, BROOKINGS COUNTY. These will all be reduced to "brookings" in the output.

It does not matter which other columns are present, as long as each row contains at least these two values. We have randomly generated some human case data that resemble the trends observed in WNV in SD. The following picture shows the expected format.



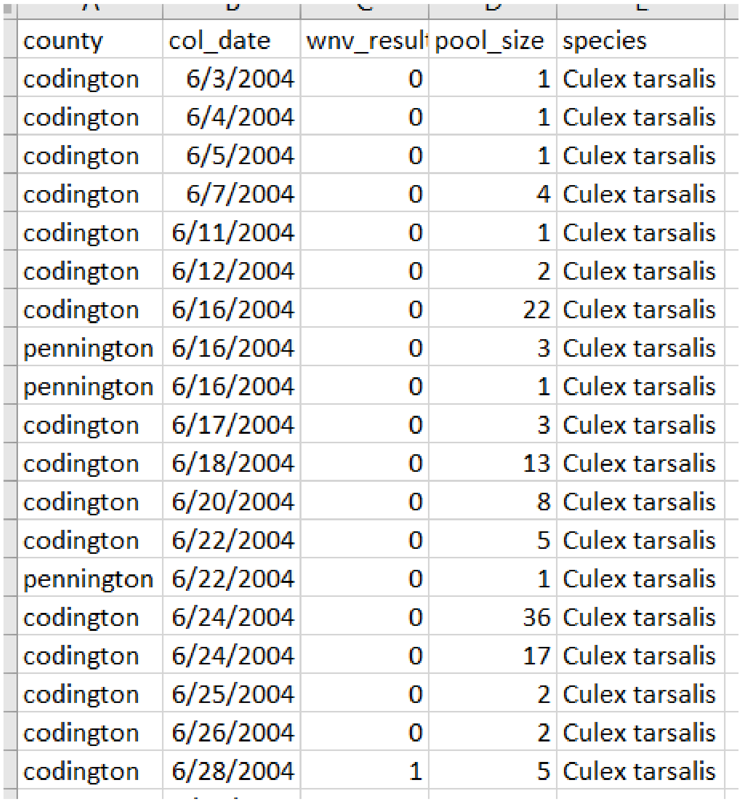
A	B
county	creationdate
brown	7/4/2004
brown	7/11/2004
brown	8/8/2004
yankton	8/29/2004
brown	6/26/2005
kingsbury	7/10/2005
lake	7/10/2005
brown	7/17/2005
brookings	7/24/2005
turner	7/24/2005
brown	7/31/2005
kingsbury	7/31/2005

Save this file in its own subdirectory. By default this is "./human case data".

Do not include any human case data from the year you are modeling. If you're making predictions for 2018, your model should be based on all human case data up to December 2017. If you include any human case data for 2018 - for example, if you know that there were a handful of cases already in mid-July 2018 and update the human case data file during the season - then the model will assume that human case data for 2018 are complete. This is almost certainly incorrect, since some cases are only reported weeks or months after diagnosis, and human case counts during the season will therefore underestimate the actual disease burden.

Formatting the vector infection data

Here we use a list of tested mosquito pools to quantify the amount of pathogen present in the environment. All tests should be present in a single CSV file, with the following columns in the following formats. The pool_size and species columns can be replaced by 1 and “default” respectively, but should be present. The wnv_result column can be 0 for a negative test or 1 for a positive test. Save this file in its own subdirectory called "./mosquito data".



county	col_date	wnv_resul	pool_size	species
codington	6/3/2004	0	1	Culex tarsalis
codington	6/4/2004	0	1	Culex tarsalis
codington	6/5/2004	0	1	Culex tarsalis
codington	6/7/2004	0	4	Culex tarsalis
codington	6/11/2004	0	1	Culex tarsalis
codington	6/12/2004	0	2	Culex tarsalis
codington	6/16/2004	0	22	Culex tarsalis
pennington	6/16/2004	0	3	Culex tarsalis
pennington	6/16/2004	0	1	Culex tarsalis
codington	6/17/2004	0	3	Culex tarsalis
codington	6/18/2004	0	13	Culex tarsalis
codington	6/20/2004	0	8	Culex tarsalis
codington	6/22/2004	0	5	Culex tarsalis
pennington	6/22/2004	0	1	Culex tarsalis
codington	6/24/2004	0	36	Culex tarsalis
codington	6/24/2004	0	17	Culex tarsalis
codington	6/25/2004	0	2	Culex tarsalis
codington	6/26/2004	0	2	Culex tarsalis
codington	6/28/2004	1	5	Culex tarsalis

Formatting the vector data stratification file.

If a stratified mosquito model is used, a spatial stratification file must be provided in its own subdirectory called "./strata". An example stratification file for South Dakota is provided with the demonstration dataset. The file must contain at least two columns: `district` contains county names and `strata` contains a unique numerical code for each stratum.

A note on standardized place names

ArboMAP was designed to model WNV on a district-week basis, but can work with any geographical unit as long as each area is assigned a unique, stable identifier. Ideally, you will work with a standard set of names (or even a numerical code like the FIPS). District names might not be not completely standardized in your sources; e.g. some sources will list "Brookings County" while others will simply use "Brookings".

ArboMAP reduces all names by deleting "County" and "Parish", removing all spaces from the district name, and putting the name in lowercase. Therefore, "BROOKINGS COUNTY" becomes "brookings" and "Charles Mix County" becomes "charlesmix". This gives the greatest chance that all places will be recognized in all data sources. However, incorrect spellings and invalid date formats (e.g. DD/MM/YYYY instead of MM/DD/YYYY) cannot be fixed by the program and will result in errors.

You will need to be sure that the shapefile you use contains a “NAME” field with human-readable names.

The typical run during the season

Updating the data during the season

Before you run the model during the WNV season, you will want to be sure you have the most up-to-date weather and mosquito data. These updates are simpler than the updates that need to be run once per year at the beginning of the WNV season, but will need to be performed every time you want to make an updated prediction.

Updating the gridMET weather data

Follow exactly the process described above, but only download data for the current year. For example, if you are modeling in 2021, start by downloading all weather data from the end of the previous WNV season to the present. The you can use the default settings to download the most recent data each week when you make a forecast. All of the new files can simply be added to the weather data folder and the software will automatically eliminate any duplicate values for format the full time series.

Updating the human case data

Do not update the human case data during the season. If you learn that there was a case last week, for example, this should not yet be included as a new line in the "human case data.csv" file. Next year, when the historical human case data are updated, this new human case will be included in the file, but during a year the "human case data.csv" file should remain untouched.

Updating the mosquito data

When new mosquito testing data are available, they need to be added to the existing mosquito data file, which must be updated and replaced in the appropriate data folder. In situations where mosquito data are managed in a database system, this can typically be accomplished via a new database query. Otherwise, the new mosquito observations will need to be appended to the mosquito data file.

Setting up the Rmd file

You will be given a file called "ArboMAP_Main_Code.Rmd", which should be placed in the working directory of your RStudio project. This is an Rmarkdown file, which does all of the calculations of modeling and produces a PDF in the end to summarize the outputs. Open RStudio and open this file.

Running the code

Once you have all the necessary data and everything is set up and formatted correctly, all you need to do is 1) open the `ArboMAP_Main_Code.Rmd` script, 2) make any necessary changes to the model parameters, and 3) execute the model run. There are two different ways to modify the model parameters. The first is to directly overwrite the parameters in the first section of the .Rmd file. To save the modified parameters, you will need to first either overwrite the existing .Rmd file or save it using a new file name. The second way to modify the parameters is to select the Knit with Parameters option from the Knit pull-down menu in RStudio. This will bring up a menu that displays all of the default parameter settings. The user can then choose to modify some of these parameters and click the Knit button in the pop-up menu. ArboMAP is then run using the modified parameters, but no changes are made to the underlying .Rmd file.

Both of these approaches to modifying the parameters can be combined to generate a modeling workflow. For example, when implementing ArboMAP in a new locations using new data, it is helpful to create a new "master" version of the .Rmd script that contains all of the appropriate defaults for the implementation. Then, when using ArboMAP operationally to generate forecasts, Knit with Parameters can be used to change the forecast date and run ArboMAP without having to modify the .Rmd file.

The time required to run ArboMAP depends on the number and type of models being used. Running a single model usually takes just a few minutes, whereas running multiple models using thin-plate splines could take 20 minutes or longer. At the end, a PDF will be generated in the same directory as the Rmd file. This PDF will always have the same name as the .Rmd script file with a .pdf extension. It will be automatically overwritten in future ArboMAP runs, so it should be saved elsewhere with a unique file name.

Interpreting the output

Weather data

Each forecasting run will generate a report containing a number of graphics to interpret. The raw and anomalized environmental data for all years will be shown, including measurements during the current year. The anomalized data are data from which the daily average has been subtracted, so the anomaly is the measurement above or below what is normal for that point in the year.

Mosquito data

The model of WNV also summarizes the mosquito infection growth rate (MIGR) for multiple strata within the state. Every winter, the virus goes into hiding. In the early season, WNV begins replicating and spreading among birds and mosquitoes. The MIGR is a measure of how quickly that's occuring. Below we show the estimates of the MIGR over time per year. If the point is above 0, then mosquito infections are growing in pools more quickly than in the average year, and more human WNV should be expected.

Human data

It will typically be the case that there are long stretches of time in which there are no human infections. Trying to model these makes most algorithms unstable - if you try to model too many 0s, you are likely to see odd results in your model fit statistics, AIC, etc. Hence, we only model cases within a certain time period - in the default, synthetic data, this is between weeks 24 and 42 of the year. This range can be tightened or relaxed by adjusting the `humancasealpha` parameter in the settings, which retains 100(1-alpha)% of the cases to model.

Multi-year WNV forecasting chart

Next are the model outputs. For each week of a year, the model estimates what proportion of districts will report at least one human case based on fits to historical data (black). Each model specified in the list at the beginning of the code will be assigned one line.

Current-year WNV forecasting chart

Next are model outputs for the year you're estimating and the two years you've selected for comparison. The solid, vertical line indicates the week of the year for which predictions are produced.

Model fit statistics

A table of fit statistical is provided for the calibration of each model. Statistics include the Akaike Information Criterion (AIC) and the area under the receiver operating characteristic curve (AUC).

Positives-to-cases chart

ArboMAP models positive district-weeks, which are either 0 if there were no cases in that district in that week, or 1 if there was at least one case. We can attempt to convert this to total case counts, by fitting a smooth curve (red, dotted) to the relationship between positive district-weeks and total case counts in a year (each dot). Predictions are summarized in the table below.

Current-week WNV absolute risk map

Next are the absolute probabilities of each district reporting at least one human case during the week in question. If the district is darkest blue, then the model says there is no probability of reporting a case during the week you've selected. This is often the case in the early or late seasons, especially in the lower-population districts. If a district is brightest red, then the model says the district will definitely report at least one human case this week. Each model will have a map.

Current-week WNV relative risk map

Next are the relative probabilities. A district in the early season may have low risk because the virus is not yet circulating, but is that lower or higher than usual, for that district, during that week of the year? For example, Brookings County SD on May 15th 2018 will have low risk, but is that risk higher than usual when compared to Brookings on May 15th in 2004-2017? If a district is red, this means that risk is predicted to be substantially higher than average in that district, during the week in question. If blue, then lower. If yellow, then average. It is important to remember that red and blue do not mean high or low absolute risk. They mean high*er* or low*er than usual.* A district that is lower than usual can still have many cases in a week; e.g. Brown County SD in early August can be lower than usual, but will still likely report cases.

The predictions of all models are averaged to produce this map.

Estimated dependence functions

Because ArboMAP allows you to run a variety of models with different forms, the output in this series of graphs may vary wildly depending on your chosen formulas. Additionally, some of the outputs may be difficult to interpret. In all cases, above each graph you will see "model: cub-sv-nonanom" or some other model name which indicates which model the dependence function comes from. Then:

- s(doy) indicates a smooth over the day of the year (doy). This is usually a cyclical term used in anomalized models, which models the general trend over all years in the sample. This is relative - the lower the function at some day of the year, the less likely you are to see human WNV cases on that day of the year, all else being equal. Generally, this should range between -5 and 5 - if it goes beyond this range, it is likely that you are modeling too many empty district-weeks and should raise your humancasealpha so that more empty weeks are cut off.
- s(lag):variable is a single distributed distributed lag function, showing the dependence of risk now (lag=0) on environmental data some time in the past. Var1/2 are whichever variables you have chosen (temperature and vapor pressure deficit) in the settings. As an example, if the estimated distributed lag is positive at lag=120, then the environmental variable four months ago correlates with an increase in human cases today. If you used anomalized environmental data, these variables will be prefixed with "anom_".
- te(lag, doymat):variable indicate how the risk today (lag=0) depends on environmental data at some point in the past. Var1/2 are whichever variables you have chosen (temperature and vapor pressure deficit) in the settings. If you used anomalized environmental data, this will have the "anom_" prefix. If you have chosen a seasonally-varying model, then these will depend on the day of the year and three different dependence functions will be shown. For example, precipitation in the winter (near doy=0) will not have much of an effect on human risk, whether two days (lag=2) or three months (lag=90) ago. If these lines are essentially all the same, you might achieve a better fit with fixed (fx) rather than seasonally-varying (sv) distributed lags.

Troubleshooting

Filenames and unrecognized escapes

The most common error will be an error in filenames. If you are told that a file is not found, it is likely that there is a typo in the file path or file name.

No estimates, blank maps

Unfortunately, if the Rmd file refuses to produce estimates (e.g. the red line is missing in the graphs or the maps are completely blank) then it's likely that a file is missing, is in an improper format, or there are extra files where they do not belong; e.g. only gridMet files should go in the gridMet data directory.

Bad estimates

This is usually because of bad mosquito infection data. You will want to carefully check the mosquito infection data to ensure they are in the correct format. You will often be able to see this on the graph of the mosquito infection growth rate - a year in the past will dip down to very low levels of infection, but this does not seem historically accurate. Another problem can occur if you enter district names incorrectly. If the data tell the program that a case occurs in a district that does not exist, then this case will not be counted, and a district's risk will be artificially low. If this happens with enough districts (e.g. if every district is followed by a space in the data file) then a whole year's risk will be 0.

Relevant scientific papers

- Chuang, T. M. B. Hildreth, M. B., D. L. VanRoekel, and M. C. Wimberly. 2011. Weather and land cover influences on mosquito populations in Sioux Falls, South Dakota. Journal of Medical Entomology 48: 669-679.

- Chuang, T., G. M Henebry, J. S. Kimball, D.L. VanRoekel-Patton, M. B. Hildreth, and M. C. Wimberly. 2012a. Satellite microwave remote sensing for environment modeling of mosquito population dynamics. Remote Sensing of Environment 125: 147-156.
- Chuang, T., C. W. Hockett, L. Kightlinger, and M. C. Wimberly. 2012b. Landscape-level spatial patterns of West Nile virus risk in the northern Great Plains. American Journal of Tropical Medicine and Hygiene 86: 724-731.
- Chuang T., and M. C. Wimberly. 2012. Remote Sensing of Climatic Anomalies and West Nile Virus
- Davis, J. K., G. P. Vincent, M. B. Hildreth, L. Kightlinger, and M. C. Wimberly. 2018. Improving the prediction of arbovirus outbreaks: a comparison of climate-driven models for West Nile virus in an endemic region of the United States. Acta Tropica 185: 242-250.
- Davis J. K., Vincent G. P., Hildreth M. B., Kightlinger L., Carlson C., and M. C. Wimberly. 2017. Integrating Environmental Monitoring and Mosquito Surveillance to Predict Vector-borne Disease: Prospective Forecasts of a West Nile Virus Outbreak. PLoS Currents Outbreaks. 2017 May 23. Edition 1. doi: 10.1371/currents.outbreaks.90e80717c4e67e1a830f17feeaaf85de.
- Hess, A., J. K. Davis, and M. C. Wimberly. 2018. Identifying environmental risk factors and mapping the distribution of West Nile virus in an endemic region of North America. GeoHealth 2:
- Wimberly, M. C., M. B. Hildreth, S. P. Boyte, E. Lindquist, and L. Kightlinger. 2008. Ecological niche of the 2003 West Nile virus epidemic in the northern Great Plains of the United States. PLoS One 3: e3744.
- Wimberly, M. C., P. Giacomo, L. Kightlinger, and M. B. Hildreth. 2013. Spatio-temporal epidemiology of human West Nile virus disease in South Dakota. International Journal of Environmental Research and Public Health 10: 5584-5602.
- Wimberly, M. C., A. Lamsal, P. Giacomo, and T. Chuang. 2014. Regional variation of climatic influences on West Nile virus outbreaks in the United States. American Journal of Tropical Medicine and Hygiene 91: 677-684.