P-MEDDS User Manual

Predictive Science Inc.

May 16, 2016

Contents

1 Introduction

P-MEDDS provides a robust set of of tools for modeling the spread of infectious diseases. Using a deterministic compartmental S-I-R model and a robust Markov-Chain-Monte-Carlo (MCMC) fitting procedure P-MEDDS can quickly characterize an incidence profile in real time providing estimates for the:

- individual level epidemic severity (as described by the proportion p_C of infections that result in clinical cases)
- and the epidemic transmissibility (measured by the basic reproduction number R_0)

Publicly available weekly influenza-like-illness (ILI) data is included in the package (from the CDC, Google Flu Trends (GFT and GFT$^+$), and the World Health Organization (WHO)) along with weekly averaged specific humidity and school opening/closing schedule. The last two sets of data are needed for three of the five different models for the time dependence of the reproduction number $R_0(t)$ that the user can choose from.

Upon completion of the MCMC fitting procedure, the **P-MEDDS** package analyzes the results and produces an extensive set of publication-quality plots (in PDF and PNG formats) and tables (in .csv format). The complete history of the run and the fitting procedure is saved as an '`RData`' file which the user can later load and further analyze.

P-MEDDS also provides a computationally efficient version of the Wallinga-Teunis (W-T) likelihood-based procedure for estimating the daily effective reproduction number using an observed epidemic curve [insert ref.]. This algorithm can be applied to the 2003 SARS epidemic data which is included in the **P-MEDDS** package. As in the case of influenza modeling, **P-MEDDS** will produce a set of publication-quality figures and tables when the estimation procedure is completed.

2 Example and Test Scripts

P-MEDDS includes an examples directory with three sample `R` drivers that demonstrate how to use the package:

- `demo.MCMC.R`: This script demonstrates the basic functionality and outputs of pkgP-MEDDS. The user specifies a flu season and data type: Military, CDC, or GFT/GFT$^+$. Parameters are automatically set for a short MCMC optimization.
- `demo.EMCEE.R`: Similar to *demo.MCMC.R*, this script demonstrates the alternate fitting routine 'emcee'.
- `demo.SYNTHETIC-DATA.R`: This demo takes several data and SIR model parameters from the user and uses these to generate epidemic data (weekly incidence). The code then initializes the MCMC routine with random parameter values. In this way it is demonstrated that the MCMC procedure converges to the correct values for data with known parameters.
- `example.driver.R`: This script can be used to model any of the ILI data in the **P-MEDDS** data base: Military, CDC or GFT/GFT$^+$. The script explains every parameter that the package requires and shows the default values for each.
- `example.interactive.R`: This is an interactive version of the example.driver.R script. It prompts the user to select each of the parameters for a **P-MEDDS** run. It explains the options and uses their defaults if the user provides an incorrect value.
- `example.wt.R`: This script demonstrates how the **P-MEDDS** package can be used to model the 2003 SARS data using the Wallinga-Teunis procedure.

For more information on each of these `R` scripts see `README.example.md` file in the `examples` directory. Inside the `examples` directory there is a sub-directory `tests.output` with sample output files for the different models supported by **P-MEDDS**. After reading this manual we suggest that the User familiarize him/her-self with the package by using these scripts and modifying them as needed.

Two codes found in the 'test' directory may also be helpful:

- `TestInstall.R`: Runs several simulations and compares the results to reference files. Ideally this will catch any major issues with the local **P-MEDDS** installation.
- `TestChanges.R`: This script is for use by developers. After making changes to the **P-MEDDS** package, `TestChanges.R` will test the new code against all/most combinations of models and data. In the case of an error, details of the error are printed to the screen. Notice: the runtime for this code is approximately 20 minutes.

3 Influenza Modeling

When modeling a specific ILI profile (either military or civilian) we consider each data set as independent and use a deterministic S-I-R compartmental model with a time dependent reproduction number $R_0(t)$:

$$\frac{dS}{dt} = -\frac{R_0(t)}{T_g}\frac{SI}{N_{total}}, \tag{1}$$

$$\frac{dI}{dt} = \frac{R_0(t)}{T_g}\frac{SI}{N_{total}} - \frac{I}{T_g}, \tag{2}$$

$$\frac{dR}{dt} = \frac{I}{T_g} \tag{3}$$

where S represents the number of susceptible individuals, I is the number of infectious individuals, R is the number of recovered individuals, and $N_{total} = S + I + R$ is the total population (see below on how we estimate it for both the civilian and military populations). The time-parameter t_0 is used to set initial conditions for the S-I-R equations as follows

$$\begin{aligned} S(t_0) &= N_{total} - 1 \ , \\ I(t_0) &= 1 \ , \\ R(t_0) &= 0 \ . \end{aligned}$$

The S-I-R equations model the total population, but the data is the number of weekly observed cases or incidence rate (I_R). The weekly incidence rate is calculated from the continuous S-I-R model by discretizing the rate-of-infection term $\frac{R_0(t)}{T_g}\frac{SI}{N_{total}}$:

$$I_R(t_i) = B + p_C \int_{t_i-1}^{t_i} \frac{R(t)}{T_g}\frac{S(t)I(t)}{N_{total}}dt \ , \tag{4}$$

scaling by percent clinical p_C, and adding a baseline B. p_C is the proportion of infectious active duty/civilian population that present themselves to a clinic with ILI-small symptoms and B is a constant number of non-SIR cases or false-ILI. The integral runs over one week determining the number of model cases for week t_i. This is how **P-MEDDS** relates its internal, continuous SIR model to the discrete ILI data. In the **ILI Methods** section we describe the procedure we use for fitting this property to the specific ILI profile.

In the case of influenza the average infectious period T_g is generally assumed to be 2.6 days, but the user can change this or even fit it (though caution should be taken as this may make the numerical procedure unstable). **P-MEDDS** enables the user to use five different time dependent models for the time dependent reproduction number, $R_0(t)$. To enable this, we write the transmission term in the most general way as a product of the basic reproduction number, R_{0min}, times the various time dependent terms:

$$R_0(t) = R_{0min} \cdot F_1(t) \cdot F_2(t) \cdot F_3(t) \tag{5}$$

The first time dependent term $F_1(t)$, allows for a dependence of the transmission rate on the specific-humidity (SH). In a series of papers [insert refs.] Shaman et al. and others [more refs] have argued that both influenza virus transmission and influenza virus survival are affected by absolute humidity which, in temperate regions, has a seasonal oscillation with a minimum in the winter and a maximum in the summer. We follow Shaman et al.[ref. here] and relate the SH to the reproduction number as:

$$F_1(t) = 1 + \Delta_R \cdot \mathrm{e}^{a \cdot q(t)} \tag{6}$$

In the above equation, and unlike the work published by others, the values of the parameters a and Δ_R are fitted. Δ_R is allowed to be positive or negative (between ± 1), and the effect of the SH term can be turned "off" by setting: $\Delta_R = 0$ and not asking **P-MEDDS** to optimize it. (In this case there is of course no need to optimize a either.) The specific humidity, $q(t)$ for all the military bases and CDC regions is provided by **P-MEDDS** using the Phase-2 of the North American Land Data Assimilation System (NLDAS-2) data base provided by NASA `http://ldas.gsfc.nasa.gov/nldas/NLDAS2model.php/`. The NLDAS-2 data base provides hourly specific humidity (measured 2-meters above the ground) for the continental US at a spatial grid of $0.125°$ which we average to daily and weekly SH. The weekly data is then interpolated for the military bases and averaged for the CDC regions. For military bases outside the continental US, and for the states of Alaska and Hawaii we obtained the SH data from NOAA's NCEP-NCAR Reanalysis project (see for example: `http://iridl.ldeo.columbia.edu/SOURCES/.NOAA/.NCEP-NCAR/.CDAS-1/.DAILY/.Diagnostic/.above_ground/.qa/`) which provides daily (again 2-meter above ground) SH data on a spatial grid of $2.5°$ for the entire world. This data is averaged and

interpolated using the same procedure as for the NLDAS-2 data set. The **P-MEDDS** SH data base is continually updated using these two data set sources.

The second time dependent term $F_2(t)$ allows for dependence of the transmission rate on school schedule. During the 2009-2010 H1N1 pandemic health officials around the world had to consider the potential benefits of reducing transmission by closing schools, against the high economic and social costs of such a drastic measure. In our formulation the transmission rate can depend on the school closure as follows:

$$F_2(t) = 1 - \alpha \cdot p(t) \tag{7}$$

where $p(t)$ is:

$$p(t) = \begin{cases} 0 & \text{if school is open} \\ 1 & \text{if school is closed} \end{cases} \tag{8}$$

When publicly available, data on schools schedule for the military bases was obtained directly from school districts that include (or are within) these bases. Unavailable data for past years was inferred from this data.

For data at the state/region/national level, state school schedules were approximated by averaging the public school schedule from the three largest cities in that state. Approximations for region schedules are determined by population-weighted average of state schedules. The same process is then applied to the regions to recover a national school schedule. Unlike the military data, here the value of $p(t)$ is allowed to vary between 0 and 1. For example in week t_i, if all schools are out for the entire week then $p(t_i) = 1$. However, if all schools have Monday and Tuesday off (missing 2 of 5 days) then $p(t_i) = 0.4$. Similarly, if 3 of 10 schools have spring break (entire week off), but the other 7 schools have a full week of class then $p(t_i) = 0.3$. And of course if all schools have a full week of class then $p(t_i) = 0$.

P-MEDDS models the effect of school closure by optimizing the parameter α which is in the range $0 - 1$. (The larger is α the more $R_0(t)$ is reduced by school closure and conversely small values of α indicate that the school schedule is not an important factor in determining the ILI profile.) As in the case of the SH term, $F_1(t)$ above, the school schedule term can be turned "off" by setting $\alpha = 0$ and not asking **P-MEDDS** to optimize it. Conversely, the user can model the joint effect of SH and school schedule by asking **P-MEDDS** to optimize all the parameters that control $F_1(t)$ and $F_2(t)$.

The third, and last, time-dependent term, $F_3(t)$, has a simple "box-like" shape and it allows the user to model an arbitrary behavior modification that can drive the transmission rate up or down for a limited period of time:

$$F_3(t) = 1 + \Delta \cdot H(t) \tag{9}$$

where

$$H(t) = \begin{cases} 1 & \text{when } t_s \leq t < t_f \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

We have found this term to be useful when modeling the military data and certain civilian data sets (e.g. San Diego County weekly number of ILI cases http://www.sandiegocounty.gov/hhsa/programs/phs/community_epidemiology/dc/influenza.html). By allowing the parameter Δ to be between -1 and 1 **P-MEDDS** can model both an increase and decrease in transmission due to behavior modification. Since this term is similar in many ways to the school closure term, $F_2(t)$, both should **not** be used at once. This term can be used together with the SH term, $F_1(t)$, though we have not explored this option yet. As in the case of SH and school schedule, this term can be turned "off" by setting $\Delta = 0$ and not asking **P-MEDDS** to optimize it.

Finally, the user can choose to use a simple S-I-R model with a constant transmission rate:

$$R(t) = R_{0min} \tag{11}$$

by setting

$$\Delta_{SH} = \alpha = \Delta = 0 \tag{12}$$

and asking **P-MEDDS** to **only** optimize the parameter R_{0min}.

Summary of Models

As described above, there are a number of ways to define $R_0(t)$. The **P-MEDDS** package allows the user to select from five different $R_0(t)$ models. Table 1 contains a brief description of each model and specifies which parameters are being optimized. Parameters that are not being optimized are generally set to their default value.

Table 1: **P-MEDDS** Models

Model #	Description	Optimized Parameters
1	School and specific humidity terms	$R_{0min}, \Delta_R, a, t_0, B, pC, \alpha$
2	Specific Humidity only	$R_{0min}, \Delta_R, a, t_0, B, pC$
3	School only	$R_{0min}, t_0, B, pC, \alpha$
4	Constant R_0	R_{0min}, t_0, B, pC
5	Stepped-R_0 (see eqs (9) & (10))	$R_{0min}, t_0, B, pC, \Delta, t_s, \Delta t$

Several example $R_0(t)$ profiles for the 2014-2015 flu season are shown in figure 1. Here, the specific humidity and school schedule corresponding to CDC region 5 is used. The figure depicts sample $R_0(t)$ profiles for models 1, 2, 3, and 5. The profile for model 4 is constant ($R_0(t) = R_{0min}$) and therefore has been omitted from the illustration. In these figures, the cyan markers denote scheduled school breaks and the height of the marker indicates the proportion of student-days missed for a given week. A marker height of 0.5 indicates that all schools are out for the entire week and lower values indicate less days/schools on break. Where all schools are in session for the entire week, no marker has been plotted. This does not take into account sick-days for individual students. The top-left illustrates a simple two-value step function. The primary value of $R_0(t)$ is R_{0min}, then for dur weeks starting at t_s, the value of $R_0(t)$ increases(decreases) to $R_{0min} * (1 + delta)$. In the top-right plot, model 3 is considered. Here $R_0(t)$ depends only on the scheduled school breaks. A $R_0(t)$ profile resulting from model 2 appears in the bottom left. The relationship between specific humidity and reproduction number is from (6). Finally, the lower-right plot shows the combined effects of school closures and specific humidity present in model 1.

A summary of parameters is found in table 2. The "Optim Range" is the range of values that the optimizer is allowed to vary over. For the parameters that control $R_0(t)$, their default value is often 0. Thus when a parameter is not being optimized, it is set to zero which "turns-off" that term. Both the default and range of values for baseline parameter B depend on B_{est}. For a single flu-year (ex. Summer 2010–Summer 2011), B_{est} is the average of the first five and last five weeks of ILI data.

3.1 ILI Data

We obtained data from the Armed Forces Health Surveillance Center (AFHSC) consisting of outpatient visits to permanent military treatment facilities (MTFs) by active duty military personnel for a range of ICD (international classification of diseases)-9 codes associated with respiratory-related illnesses between January 1, 2009 and April 30, 2011. For each record, the data contained: a unique study identifier for the individual; ICD-9 codes associated with that visit; and the zip code (5 digits) of the clinic location. We used the zip code of the reporting clinic

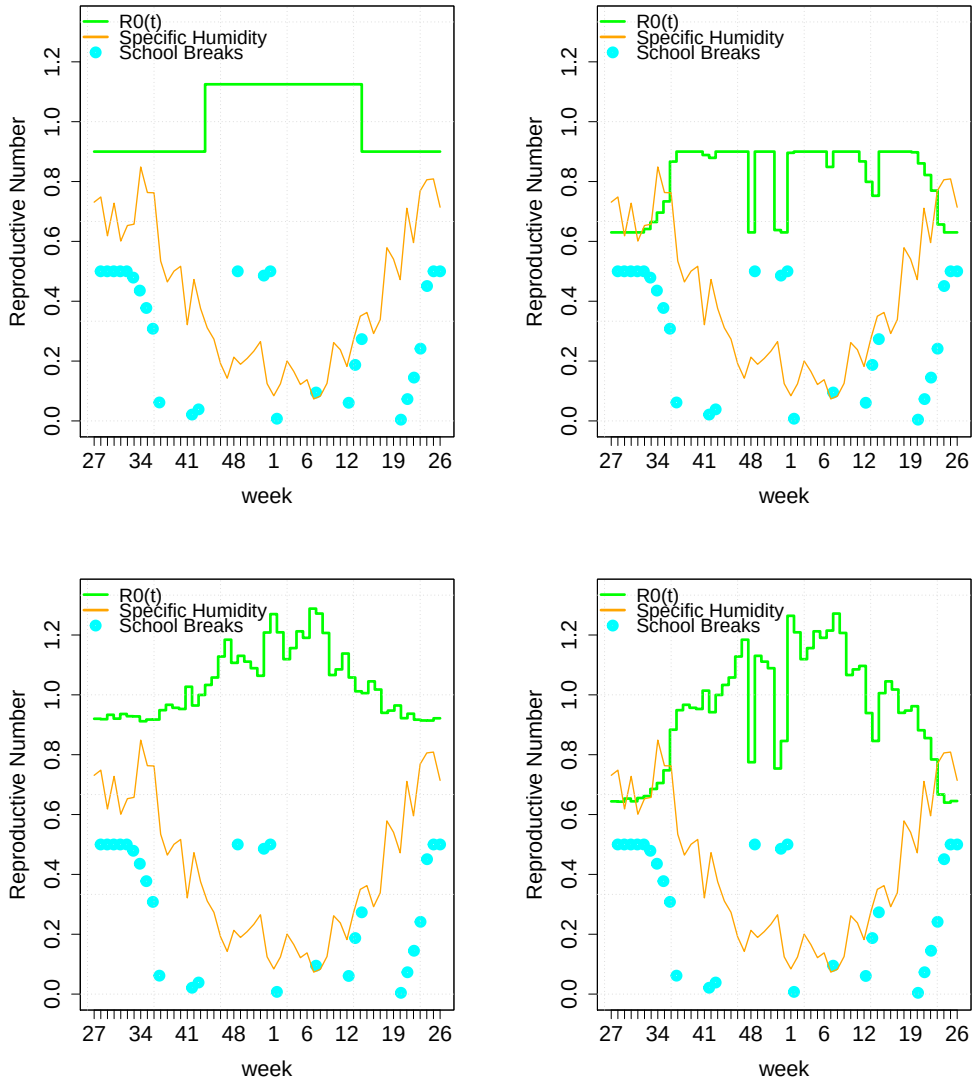

Figure 1: $R_0(t)$ profiles for the model parameters: $R_{0min} = 0.9, \Delta_R = 0.6, a = 300, \alpha = 0.3, t_s = 18, dur = 23, \Delta = 0.25$. (Top Left) model 5: stepped-$R_0(t)$. (Top Right) model 3: school only. (Bottom Left) model 2: specific humidity only. (Bottom Right) model 1: school and specific humidity terms.

as a proxy with which to define a military installation: we do not explicitly represent military installations or bases, rather, we assume that the case reports from the same zip code are from the same military installation. Each record (an anonymized Study ID) was assigned as either "ILI-large" ($n = 1,336,471$) or "ILI-small" ($n = 27,582$) using a set of classifications based on ICD-9 codes [ref here]. The definition of ILI-large was broader and included non-specific diagnosis such as: 'viral infection' and 'acute nasopharyngitis'. The definition of ILI-small was more constrained and included: 'influenza w/other respiratory manifestations' ($n = 25,293$), 'influenza with manifestation not elsewhere classified (NEC)' ($n = 1,006$), ' infectious upper respiratory, multiple sites, acute NEC' ($n = 897$), and 'influenza with pneumonia' ($n = 404$).

Table 2: **P-MEDDS** Model Parameters

Parameter	Description	Default	Optim Range
R_{0min}	Baseline reproduction number	1.4	0.5–4.0
T_g	Average infectious period (weeks)	2.6/7	(1/7)–1
t_0	Infection start week ($I(t_0) = 1$)	1.0	1.0–40.0
B	Baseline: mag. of non-SIR ILI data	B_{est}	$(0.1 * B_{est})$–$(10 * B_{est})$
pC	Percent clinical	0.01	10^{-6}–1.0
α	Reduction in R_0 for school closed	0	10^{-6}–1.0
Δ_R	Max change in R_0 from specific humidity	0	10^{-3}–2.0
a	Exponential for specific humidity R_0-term	200	1.0–500.0
Δ	Increase/Decrease in R_0 (model 5)	0	-1.0–1.0
t_s	Start time of change in R_0 (model 5)	0	1.0–40.0
dur	time duration: $t_f = t_s + dur$ (model 5)	0	1.0–40.0

We further trimmed the data temporally to cover the period from April 1, 2009 through June, 1, 2010 which is the 2009 H1N1 pandemic period. We ranked the military installations by size according to the total number of ILI-small cases they reported.

In the **P-MEDDS** data-base we provide the data for the top-50 largest profiles, 47 of which, were located within the USA. Of the remaining three, one was located in Landstuhl, Germany, and two were located in Japan (Misawa and Yokosuka).

For each of these 50 military installations we provide our estimate of the total population (N_{total}), the "denominator data". Our method for estimating these sizes relied on the use of the total number of visits to a clinic for all causes as a proxy for the total number of active duty personnel at that location [ref]. The coefficient of proportionality Ω was estimated by using a subset of the installations for which reasonably reliable estimates for the total population have been published.

We obtained civilian data through a variety of means. County- and State-level data were generally acquired directly from the appropriate public health services department.

The CDC Influenza-like Illness Surveillance Network (ILINet) Human and Health Services (HHS) region and national data were downloaded from the CDC hosted web application FluView http://gis.cdc.gov/grasp/fluview/fluportaldashboard.html for all available dates (Week 40, October 10th, 1997 through Week 45, November 9th, 2014).

The CDC Patient ILINet consists of more than 2,900 outpatient healthcare providers in all 50 states, Puerto Rico, the District of Columbia and the U.S. Virgin Islands reporting more than 30 million patient visits each year. Each week, approximately 2,000 outpatient healthcare providers around the country report data to CDC on the total number of patients seen and the number of those patients with influenza-like illness (ILI) by age group (0-4 years, 5-24 years, 25-49 years, 50-64 years, and >= 65 years). For this system, ILI is defined as fever (temperature of 100F or greater) and a cough and/or a sore throat without a KNOWN cause other than influenza. Sites with electronic health records use an equivalent definition as determined by public health authorities. The CDC Influenza surveillance data collection is based on a reporting week that starts on Sunday and ends on Saturday of each week. With the exception of the 2009 influenza pandemic, each flu season starts on Week 27 and ends on Week 26 of the following year. For 2009, the flu season started Week 13, 2009 (3-30-2009) and ended Week 26, 2010 (6-28-2010). For more information on ILINet see: http://www.cdc.gov/flu/weekly/overview.htm.

Because **P-MEDDS** requires an absolute number of incidences per week, the CDC ILINet data must be converted from percent ILI cases per patient to ILI cases. We estimate the absolute number of weekly ILI cases by dividing the weighted number of ILI cases in the CDC data ("X.WEIGHTED.ILI" column) by 100 and multiplying it by the total weekly number of patients ("TOTAL.PATIENTS" column).

Google Flu Trends data (at the national, state and city level) was obtained directly from the

GFT web site: http://www.google.org/flutrends/us/data.txt. GFT attempts to quantify ILI cases through a proprietary model using key terms from search queries (originally discussed in Ginsberg et al., 2009; see also Cook, 2011 and Copeland, 2013). GFT ILI estimates were calculated from Week 39, September 28th, 2003 to Week 45, November 9th, 2014. The GFT data was aggregated to the same ten HHS regions as for the CDC data and the flu seasons were also defined using the same dates. In addition to the original GFT data, **P-MEDDS** now includes the 2014 model revision update, referred to as: GFT$^+$, see: https://www.google.org/flutrends/historic/us-historic-v3.txt. The GFT$^+$ data set is limited to the national and HHS regions level. For more information on GFT and GFT$^+$, see: http://www.google.org/flutrends/about/how.html.

Denominator civilian data was obtained from the US 2010 census data center: http://www.census.gov/2010census/data/. The estimated inter-censal state and national population sizes between 2000 and 2010 were obtained from http://www.census.gov/popest/data/intercensal/state/state2010.html and http://www.census.gov/popest/data/intercensal/state/tables/ST-EST00INT-01.csv.

The projected state and national population sizes from 2010-2013 were obtained from http://www.census.gov/popest/data/datasets.html and http://www.census.gov/popest/data/national/totals/2013/files/NST_EST2013_ALLDATA.csv.

Each region population size was calculated by summing the population sizes of every state in the region.

Week Indexing

All of the ILI data in **P-MEDDS** is in the form of weekly reporting. To be consistent with week numbering and dates, we have adopted the CDC calendar. This means that a week begins on Sunday and ends on Saturday. For the purpose of numbering the weeks of a year this means that week number 1 is the week containing the first Wednesday of the year. Any January days occurring before week 1 are considered part of the final week (52 or 53) of the previous year. ref{MMWR}

Data Selection Variables

The function *get.data()* provides easy access to all the data types in **P-MEDDS**. Below is a list of descriptions for the input variables of *get.data()*.

dataType - character string specifying data type. "MPZ" is military data by zip code. "CDC" is Influenza-Like-Illness (ILI) data from the Centers for Disease Control by region or national. "GFT" and "GFTPlus" are two versions of Google Flu Trends. The GFT datasets are an estimation of CDC ILI and thus are divided into the same regions as the CDC data.

national - a boolean flag that determines national or regional data for data types: "CDC", "GFT", and "GFTPlus". TRUE = national, FALSE = regional.

iregion - for the case of dataType = "CDC" || "GFT" || "GFTPlus" and national=FALSE, this integer specifies which CDC region (1-10).

myMPZ - for dataType = "MPZ", this variable specifies the base zip code.

job.year - integer specifying the flu season *start* year. Typically, this is the only time-variable required, however the user may customize the time-window by setting the variables: 'week.start', 'end.year', and 'week.end'. The week numbers entered for 'week.start' and 'week.end' are interpreted by the CDC calendar convention explained above.

wflag - integer that indicates which type of weekly weighting is to be used. The default is $wflag = 0$; equal weights for all weeks.

wweight - for $wflag = 1$, this is the user-supplied weights. A floating-point vector with the same number of elements as the epi data (ex. mydata$cases).

3.2 ILI Methods

The **P-MEDDS** ILI fitting procedure determines the joint posterior distribution for the model parameters (these are chosen by the user as explained in the previous section) using a Metropolis-Hastings Markov Chain Monte Carlo (MCMC) procedure. It should be stressed that although we often refer to the MCMC algorithm in the context of an optimizer, it is better characterized as a probability distribution mapping routine. As such, the random walk will likely spend the majority of its time in the neighborhood of the optimal (most likely) solution. However it's purpose is to find a distribution of most-likely solutions, not necessarily *the* most likely solution.

After the user sets the details of the model (that is which parameters to optimize and which not and what value to set the later to) the user needs to tell **P-MEDDS** how many MCMC chains to run and how many steps to take in each chain. **P-MEDDS** will then randomly initialize the parameters that are optimized (using a log-uniform distribution for all the parameters except the one that can be negative) integrate the coupled S-I-R and influenza incidence equations and generate a candidate ILI profile. The likelihood of this solution is calculated using the Akaike Information Criterion (AIC), which is a measure of the relative goodness of fit of a model:

$$\mathrm{AIC} = -2\log(L(\hat{\theta}|I_C)) + 2K \tag{13}$$

where $\log(L(\hat{\theta}|data))$ is the value of the maximized log-likelihood over the model parameters (θ), given the observed cases I_C. When the total number of parameters (K) is large relative to the sample size (n), the reduced Akaike Information Criterion is preferred:

$$\mathrm{AIC}_c = -2\log(L(\hat{\theta})) + 2K + \frac{2K(K+1)}{(n-K-1)} \tag{14}$$

and this is what **P-MEDDS** uses. The log-likelihood stems from a Poisson probability density

$$\log(L(\theta|I_C)) = \sum_i w_i \Big(I_C(t_i) \cdot \log\left(I_R(t_i,\theta)\right) - I_R(t_i,\theta) - \log\left(I_C(t_i)!\right) \Big) , \tag{15}$$

where $I_R(t_i,\theta)$ is the model point generated for week t_i as a result of parameter set θ. Additionally, the vector w has been added to allow the user to vary the weight given to each weekly data point. To maximize $\log(L(\theta|I_C))$, the value of this likelihood is compared to a new likelihood calculated using a set of randomly displaced parameters in a standard rejection method to determine if the move is accepted or rejected. This MCMC procedure is executed as many times as the user has defined (this is the chain's length mentioned above) and **P-MEDDS** keeps the history of the chain parameters and AIC_c values. Once a chain is completed, it's history statistics and results are summarized and written to tables (csv format), a binary '`RData`' file and pdf/png plots. The **P-MEDDS** MCMC chains have a typical acceptance rate of $20\% - 60\%$. If this is not the case the user should adjust the "step-size" in the MCMC procedure. The detailed output provided by **P-MEDDS** enables the user to:

- Quickly look at the plots and evaluate the procedure/results.
- Generate detailed reports using the plots and tables prepared by **P-MEDDS**.
- Use the history of the MCMC chain to calculate any additional statistics/properties.

Alternate Ensemble Procedure *emcee*

emcee is an ensemble-based based MCMC-type procedure with shorter autocorrelation times than standard MCMC algorithms. A randomly-coupled system of walkers produces a result similar to the single walker in MCMC.

Forecasting Mode

P-MEDDS can be set to only fit the first $nweeksfit$ weeks of a flu season. When $nweeksfit$ is less than the number of weeks in the flu season, the resulting plot will show the profile fit as well as the 'predicted' ILI values for the remainder of the season. In this way **P-MEDDS** can be used to create an ILI forecast. For historic years, the actual ILI is plotted alongside the prediction curves allowing the user to gauge the accuracy of the forecast.

Numeric Parameters

reals - number of MCMC realizations (integer). Generally, each realization is an MCMC optimization-chain with a unique set of initial parameter values. The default behavior for **P-MEDDS** is to randomly select initial values for the parameters-to-be-optimized using a uniform distribution over the range of values listed in table 2. For a quick run, set *reals* to 1. For many data/model combinations, setting *reals* to 5 will be enough to find the global minimum. Depending on the number of parameters being optimized, the prevalence of 'deep' local minima in the optimization-objective function, and the number of MCMC steps, a much larger value of *reals* may be required to find the global minimum.

nMCMC - number of 'random-walk' steps per MCMC realization (integer). For a quick run, set this to 10^4. For most of the model/data combinations in **P-MEDDS**, 10^6 steps will be sufficient. However it is recommended to check the resulting parameter time-series plots for convergence. If any parameters are still trending up or down at the end of the chain, a larger value of $nMCMC$ is likely needed.

walkers - specifies the number of random-walkers for use with the EMCEE algorithm. This number should be as large as possible without reducing performance. Minimum of hundreds, recommended 1000+.

nlines - number of MCMC steps to be saved for the purpose of statistical analyses (integer). Must be less than or equal to $nMCMC$. Default value is 10^4. For shorter runs, can be set lower.

nweeksfit - number of weeks at the beginning of the dataset to fit. Must be $\leq$ than the number of weeks of data. By setting $nweeksfit$ to be less than the number of data points, the user can compare the predicted ILI to the actual ILI for weeks after $nweeksfit$.

Other Options

optTg - a boolean flag that determines if T_g is a fixed or optimized variable. TRUE: T_g is optimized; FALSE: (default) T_g is treated as a user-specified constant. It is not recommended to optimize R_0 and T_g at the same time. This may cause the optimization routine to become unstable.
iseed - seed for random number generation. For reproducible results, set *iseed* to a fixed value. If *iseed* is set to NULL or unspecified, a pseudo-random seed is generated.

3.3 ILI Results

To demonstrate the results of an ILI run with **P-MEDDS**, we will use the military data set from the 2009-2010 pandemic year and select the military zip code to be that of the first base = 23708. In terms of setting the parameters in an `R` driver for the code this implies:

```
> require(pmedds.core)
> dataType= 'MPZ'
> job.year = 2009
> myMPZ = 23708
```

Please note that we do not expect the user to know the zip codes of the military bases, **P-MEDDS** can provide it using the following statement:

```
>get.mpz.names()
```

We will run a single MCMC chain with $1e6$ step and save the history of $1e4$ of these steps:

```
> reals = 1
> nMCMC = 1e6
> nlines = 1e4
```

The infectious period, T_g, will be set to 2.6 days:

```
> Tg = 2.6
```

and we will use the model with an arbitrary "box-like" time-dependent term for the transmission term, $F_3(t)$ above. In the **P-MEDDS** code this is model number 5:

```
> imodel = 5
```

We will give the run a name and set the seed for the random number generator:

```
> job.name='example.pmedds'
> iseed = 123456
```

Set the logical parameters "debug" , "verbose" and "plot":

```
> debug = FALSE
> verbose = FALSE
> plot = TRUE
```

and the plotting device:

```
> device = 'pdf'
```

Finally, although this is a run of military data we will set the CDC region number and the logical parameter "national" (which triggers modeling of either a CDC, GFT or GFT$^+$ ILI profile of the entire US) to their default values:

```
> iregion = 1
> national = FALSE
```

All of these settings can be found with a detailed explanation in the `examples` sub-directory: the `example.driver.R` script. The best way to execute this script is either from the command line (after changing to the correct `examples` sub-directory)using:

```
> Rscript example.driver.R
```

or from within an R session using:

```
> source('example.driver.R')
```

Continuing with our explanation we now load all the data available for this military base using the 'get.data' call:

```
  mydata = get.data(dataType=dataType,myMPZ=myMPZ,iregion=iregion,national=national,
  job.year=job.year)
```

To see what the object "'mydata" contains use any of these statements:

```
 > names(mydata)
 > print(mydata)
```

Note that in order to load the data only the variables that tell **P-MEDDS** which data type for which year and base/region/national to load needs to be set. This enables the user to view the data and then decide which model to use for the transmission term, $R(t)$.

The actual call to the **P-MEDDS** engine uses all the paramters we have set above:

```
> out <- runPMEDDS(dataType=dataType,mydata=mydata,imodel=imodel,Tg=Tg,nMCMC=nMCMC,
nlines=nlines,reals=reals,iseed=iseed,debug=debug,verbose=verbose,plot=plot,device=device,
job.name=job.name,job.year=job.year)
```

As noted above all what we have executed so far can be found in the `example.driver.R` script, which is in the `examples` sub-directory. Depending on the number of MCMC chains that we chose to run and the number of steps in each chain, **P-MEDDS** will take seconds/minutes or hours to run. Once runPMEDDS is called, detailed information about the run will be written to the screen. When the job is completed some of this information can also be found in the "log.txt" file which is shown below for our short and simple example run:

```
pkg{P-MEDDS} Package Version 001

 Job Name:  example.pmedds Job Started on:  Wed Jan 07 12:36:32 2015
 Job Running on:  mmac
 OS Information:  Darwin release 13.3.0 machine x86_64
 Job Running by User:  michal
 Job Running in Directory:  /Users/michal/work/LEPR01/p-medds/examples
 All Data and Plots Saved in Sub-Directory:  /Users/michal/work/LEPR01/p-medds/examples/output
 Modeling FY  2009 - 2010
 Modeling  MPZ Data
 Modeling Base Number:  1
 Using a Fixed Base Population of:  15842
 Modeling MPZ:  23708
 Using a Fixed Generation Time of:  2.6  days
 Optimizing the Following Parameters:  R0min Baseline pC t0 delta ts dur
 Running  1  MCMC Chains
 Model Set to Number  5
 Number of MCMC Steps in Each Chain Set to  1e+06
 DEBUG set to  FALSE

 For a PDF Plot of Base Profile, Specific Humidity and School Closure See:  output/MPZ_2009_237

 For a PNG Plot of Base Profile, Specific Humidity and School Closure See:  output/MPZ_2009_237

 Running MCMC Chain Number  1

 Writing MCMC Statistics for this Chain to File:  output/param-stats-MPZ_2009_23708-1.csv

 Writing MCMC Quantiles for this Chain to File:  output/param-quantiles-MPZ_2009_23708-1.csv

 Writing MCMC Condensed Statistics for this Chain to File:  output/param-table-MPZ_2009_23708-1

Acceptance rate for Chain:  40.36

 Writing R object Data file for this Chain:  output/mcmc-MPZ_2009_23708-1.RData

 PDF Plot of MCMC Chain written to: output/chain-MPZ_2009_23708-1.pdf
```

```
 PNG Plot of MCMC Chain written to: output/chain-MPZ_2009_23708-1.png


 PDF Plot of MCMC Profiles written to: output/plot-MPZ_2009_23708.pdf

 PNG Plot of MCMC Profiles written to: output/plot-MPZ_2009_23708.png

 PDF Plot of MCMC Cumulative Attack Rate written to:output/car-MPZ_2009_23708.pdf

 PNG Plot of MCMC Cumulative Attack Rate written to: output/car-MPZ_2009_23708.png

 Elapsed Time:
user    system   elapsed
608.237   0.463 609.035
```

As shown above, **P-MEDDS** saves all the plots and tables in a sub-directory called `output`. The names of the plot files and tables includes all the information that defines the ILI profile we are modeling: type (MPZ), season (2009, starting year is indicated), zip code (23708, or region number or national in the case of civilian data). For the tables, information of each MCMC chain is written to a separate file and hence the file name includes the chain number as the last number before the file name extension. The first plot made by **P-MEDDS**, Figure 2, shows all the data for this military base: the ILI profile (red line and left y-axis), the specific humidity (blue line and right y-axis) and the school schedule (cyan bars indicate that the school is closed). This figure includes only input from the **P-MEDDS** data base and in the `tests.output` directory there are both png and pdf versions of it.

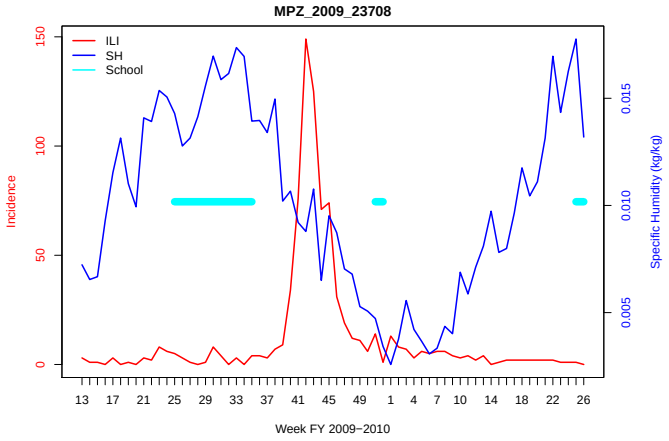

Figure 2: The ILI profile (red line), specific humidity (blue line) and school closure (cyan bars) for MPZ 23708 for the 2009-2010 pandemic year.

The results of the run are shown in Figure 3. The red line is again the ILI profile we are trying to fit. In blue we show our best estimate for it and the grey lines are 100 randomly selected estimates from the second-half of the MCMC chain. The green lines (and right y-axis show the transmission term, $R(t)$, with dark-green being our best estimate for it.

The resulting cumulative attack rate is shown in Figure 4 using the same color codes: red for data, blue for our best estimate and grey for the 100 randomly selected estimates. Finally, the history of the parameters that were optimized, and of the AIC_c score are shown in Figure 5.

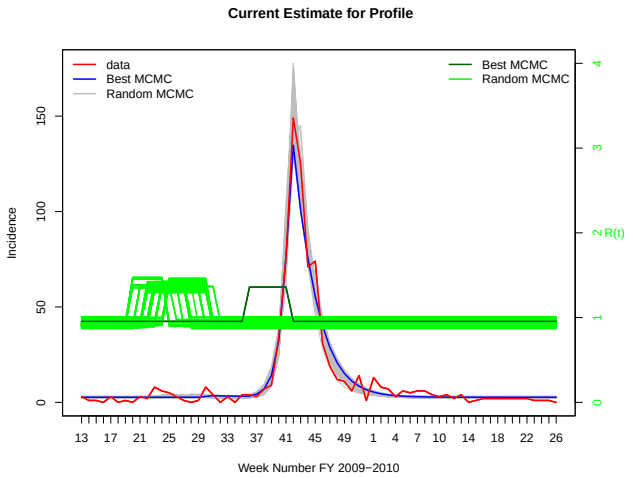

Figure 3: The ILI profile (red line), our best fit to it (blue line) and 100 randomly selected results from the MCMC chain (grey lines). The transmission term is shown in green with dark green being our best estimate for it.

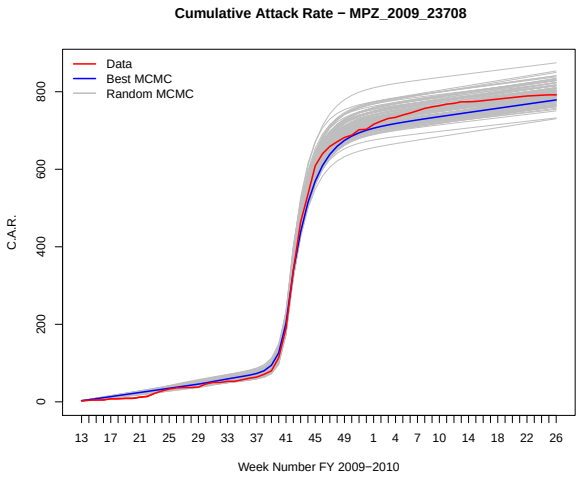

Figure 4: The ILI profile (red line), our best fit to it (blue line) and 100 randomly selected results from the MCMC chain (grey lines).

Three tables are produced by the run summarizing the statistics of each MCMC chain.

- param-stats-MPZ_2009_23708-1.csv includes the mean and standard deviation of all the parameters that **P-MEDDS** supports. Parameters that were not included in the calculation will also be shown (typically they will be set to zero) as well as parameters that were included but not optimized (e.g. the denominator data, N_{total}, and the infectious period, T_g).
- param-quantiles-MPZ_2009_23708-1.csv includes the quantiles $5\%, 25\%, 50\%, 75\%$ and 95% of all the parameters that **P-MEDDS** supports. Parameters that were not included

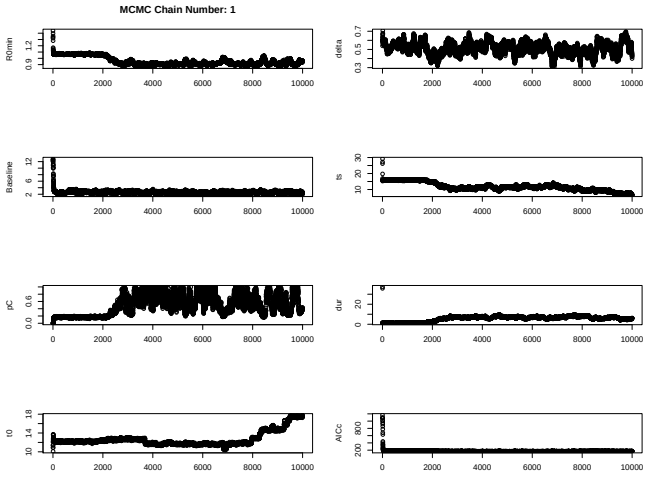

Figure 5: The history of the MCMC chain for the parameters that were optimized in our model run (four left panels and the three top right panels) along with the AIC_c score of the chain (bottom right panel).

in the calculation will also be shown (typically they will be set to zero) as well as parameters that were included but not optimized (e.g. the denominator data, N_{total}, and the infectious period, T_g).

- param-table-MPZ_2009_23708-1.csv includes the mean, standard deviation and quantiles $5\%, 25\%, 50\%, 75\%$ and 95% only for the parameters that were optimized when model number 5 is used. For model number 5 the two values that the transmission coefficient has are also calculated and reported as R_A and R_B along with their maximum denoted as R_0. This file is a summary of the above two files for the parameters that were optimized. A snapshot of the file is shown in Figure 6.

For each MCMC chain, all of the information on the run is saved by **P-MEDDS** in a binary '`RData`' file (mcmc-MPZ_2009_23708-1.RData in our example) which the user can load and reuse. The information about the the data set and all the run parameters is contained in the list object "model" and all the information about the MCMC chain is contained in the "mcmc" object which is of class "mcmc". The MCMC chain can be reused to produce any required statistics - or as an initial guess for a new optimizaition procedure. To see what was saved the user may start with:

```
> load(file='output/mcmc-MPZ_2009_23708-1.RData')
> names(model)
> summary(mcmc)
```

This completes our review of the results of a **P-MEDDS** run. We suggest first using and then modifying the scripts provided in the `examples` directory as the best way for the user to become more familiar with the requirements and capabilities of **P-MEDDS**.

4 SARS Modeling

P-MEDDS provides the data and tools to model the 2003 Severe Acute Respiratory Syndrome (SARS) outbreak using the original Wallinga and Teunis procedure [reference WallingaAndTeunis04a.pdf].

	Mean	SD	Naive SE	Time-series SE
R0min	0.92501533	0.04029575	0.00056981	0.0091589
deltaR	0	0	0	0
aparam	0	0	0	0
pC	0.64096502	0.23817483	0.00336796	0.03870204
Baseline	2.65965461	0.26158332	0.00369898	0.01698953
Tg	0.37142857	0	0	0
N	15842.0111	0	0	0
t0	23.3945539	0.28100331	0.00397359	0.0608045
alpha	0	0	0	0
delta	0.59810872	0.04895237	0.00069222	0.00801592
ts	2.56172651	0.32414819	0.00458369	0.06118758
dur	3.92987034	0.40532192	0.00573154	0.05118347
AICc	371.156627	3.51237774	0.04966756	0.28496265
RA	0.92501533	0.04029575	0.00056981	0.0091589
RB	1.47650978	0.02872062	0.00040613	0.00409326
R0	1.47650978	0.02872062	0.00040613	0.00409326

Figure 6: A snapshot of the param-table-MPZ_2009_23708-1.csv which includes the mean and standard deviation for the parameters that were optimized in our example run. See text for more details

4.1 SARS Methods

Given a time series of the number of cases by date of symptom onset the W-T method provides a numerical procedure for estimating the daily effective reproduction number, R_e. The effective reproduction number, R_e, is typically lower than the basic reproduction number, R_0, since it reflects the effects of control measures (such as school closure) and the depletion of the susceptible population, on the spread of an epidemic. The exact calculation of the daily value of R_e requires detailed knowledge about the entire course of the epidemic spread at the level of "who infected whom". In most cases this information is not available, and only the epidemic curve is observed. Likelihood based methods do provide the numerical procedure for inferring "who infected whom", using the observed dates of symptom onset, but they quickly become unfeasible since they require the calculation of all possible infection networks. Wallinga and Teunis reduce this exponential computational burden, while still retaining the likelihood-based method aspect, by replacing the calculation of the complete infection network by only pairs of cases. Assuming a known distribution of the generation interval, $w(\tau)$, the relative likelihood (p_{ij}) that case i has been infected by case j, given their difference in symptom onset time: $t_i - t_j$ is given by the ratio of the likelihood that case i has been infected by case j normalized by the likelihood that case i has been infected by *any* other case k:

$$p_{ij} = w(t_i - t_j) / \sum_{i \neq k} w(t_i - t_k), \tag{16}$$

The effective reproduction number for case j is given by the sum over all cases i:

$$R_j = \sum_i p_{ij} \tag{17}$$

and the average daily reproduction number, is the mean over R_j for all cases j with symptom onset on day t. For the 2003 SARS outbreak, Wallinga and Teunis used a Weibull distribution, with shape and scale parameters corresponding to a generation interval of 8.4 days and a standard deviation of 3.8 days, to describe the generation interval distribution, $w(\tau)$. (For more details on the W-T procedure see their original manuscript, [WallingaAndTeunis04a.pdf].)

P-MEDDS provides a numerically efficient implementation of this approximate procedure along with publicly available data for the 2003 SARS outbreak.

4.2 SARS Data

The first known case of SARS occurred in a province in southern China on November 16, 2002 and from there the infection spread to Hong Kong (late February 2003) and then by airline travel throughout the world. During 2003 outbreaks were reported in Vietnam, Singapore and Canada (in addition to China and Hong Kong). Via private communication we have obtained the SARS data for Canada, Hong Kong, Singapore and Vietnam from Prof. J. Wallinga. Using this data we were able to test the **P-MEDDS** implementation of the algorithm and ensure that it reproduces the published results for these four countries (Figure 1 in WallingaAndTeunis04a.pdf). Data for China and Canada was obtained by digitizing published figures: http://www.who.int/csr/sars/epicurve/epiindex/en/index3.html and http://www.who.int/csr/sars/epicurve/epiindex/en/index6.html, respectively. The data for China, unlike all other countries, is by date of report (and not onset). For all countries the data is in the form of a double column text file with day number and number of cases (with onset or report on that day) in each row. The code assumes that the list of (day,case) pairs is unordered and orders it as a proper time series. **P-MEDDS** can be used to model SARS data from other countries (or from the entire world, see http://www.who.int/csr/sars/epicurve/epiindex/en/ for a complete list of available WHO SARS data) as long as it is provided in this simple format.

4.3 SARS Results

We will describe the **P-MEDDS** results for a SARS run using the data from Hong-Kong and the `example.wt.R` script (in the `tests.output` directory as a guideline). Just as in the case of an ILI run we start by loading the package, giving the job a name and setting the debug, verbose and device parameters:

```
> require(pmedds.core)
> job.name="test.pmedds-wt"
> debug=FALSE
> verbose=FALSE
> device="pdf"
```

We next seed the random number generator, and set the shape (β) and scale (α) parameters of the Weibull distribution to 3.8 and 8.4 days. (As explained above this corresponds to a generation time of 8.4 days with a standard deviation of 3.8 days.)

```
> iseed = 123456
> alpha = 8.4
> beta = 3.8
```

The execution time of the code scales with the square of the cumulative number of cases ($ncase$) and the number of realizations ($ireal$): $ncase^2 \times ireal$. Setting the number of realizations to 10^4 will give numerically converged results, while still executing in real time for all countries but China which has the largest cumulative number of cases. (Specific execution times are: Singapore - 9 sec., Hong-Kong - 7 min., Canada - 20 sec, Vietnam - 2 sec., China - 100 min.). Once we select the country we want to model:

```
> mycountry = "Hong-Kong"
```

we can load the SARS data of the country:

```
> mydata = get.wt.data(mycountry=mycounrty)
```

and call the **P-MEDDS** engine (using all the parameters we set above) that will run the W-T procedure and produce all the tables and plots:

```
> out = runPMEDDS.WT((mydata=mydata,ireal=ireal,shape=beta,scale=alpha,iseed=iseed,
debug=debug,verbose=verbose,device=device,job.name=job.name)
```

As noted above all of what we have executed so far can be found in the `example.wt.R` script, which is in the `examples` sub-directory. For our choice of country to model, Hong-Kong, the run will take about seven minutes to complete. Once runPMEDDS.WT is called, detailed information about the run will be written to the screen. When the job is completed most of this information can also be found in the "log-wt.txt" file which is shown below form our example Hong-Kong run:

```
P-MEDDS Package Version 001
 Job Name:
 Job Started on:  Tue Mar 17 14:52:09 2015
 Job Running on:  mmac
 OS Information:  Darwin release 13.3.0 machine x86_64
 Job Running by User:  michal
 Job Running in Directory:  /Users/michal/work/LEPR01/p-medds/examples
 All Data and Plots Saved in Sub-Directory:  /Users/michal/work/LEPR01/p-medds/examples/output

 Modeling  Hong-Kong 2003 SARS Data
 Running  10000  Realizations for a Total of  1702  cases

 Created /Users/michal/work/LEPR01/p-medds/examples/output Directory for all the W-T Data

 Writing R object Data file:  /Users/michal/work/LEPR01/p-medds/examples/output/Hong-Kong.RData

 Writing the sampled results to file: /Users/michal/work/LEPR01/p-medds/examples/output/Rl.Hong

 Writing ordered results to file: /Users/michal/work/LEPR01/p-medds/examples/output/Rstats-Hong

 Plotting Results to: /Users/michal/work/LEPR01/p-medds/examples/output/W_T_Hong-Kong.pdf

 Elapsed Time:
user    system   elapsed
425.134   0.916   427.291
```

As shown above **P-MEDDS** saves all the plots, tables and R data files in a sub-directory called `output` with unique file names that include the country name. The plot file produced by **P-MEDDS** for this run is shown in Figure 7. The top panel is the input time series we are modeling: the probable number of SARS cases in Hong-Kong by symptom onset day from February 14, 2002 to May 20, 2003. The lower panel shows the W-T estimate of the daily reproduction number. The red and green circles are the mean and median values, respectively, and the blue bars mark the 95% CI. The grey horizontal line marks the value of $R = 1$, above this value the epidemic will spread and below it, it is contained.

The entire history of the 10^4 realizations is written in the *Rl.Hong-Kong.csv* file-a column for each realization. The user can load this file and use it for any post-run calculations. The statistics for these realizations are summarized in the *Rstats.Hong-Kong.csv* file which includes the daily statistics from these realizations (used in Figure 7): the date, day number, number of

cases, mean, median and 95% CI for the daily reproduction number. Finally, the entire input and output of the run is saved in a binary RData file: *Hong-Kong.RData*. This file can be loaded using:

```
> load("output/Hong-Kong.RData")
```

This loads an R list object `results`. To see what this list containes use:

```
> names(results)
```

The `results$mydata` has the case number, onset (or report) day, total number of cases and country name. The mean, median and 95% CI of the daily ordered reproduction number are in: `results$Rlm.order`, `results$Rlmd.order` and `results$Rlq.order`. The number of realizations that we run (10^4 in this case) is in `results$ireal` and the actual realizations are in `results$Rl`.

This completes our description of the W-T SARS modeling capabilities of **P-MEDDS**. We recommend that the User try the provided script `example.wt.R` in the `examples` directory and use it as a starting point for his/her own needs.

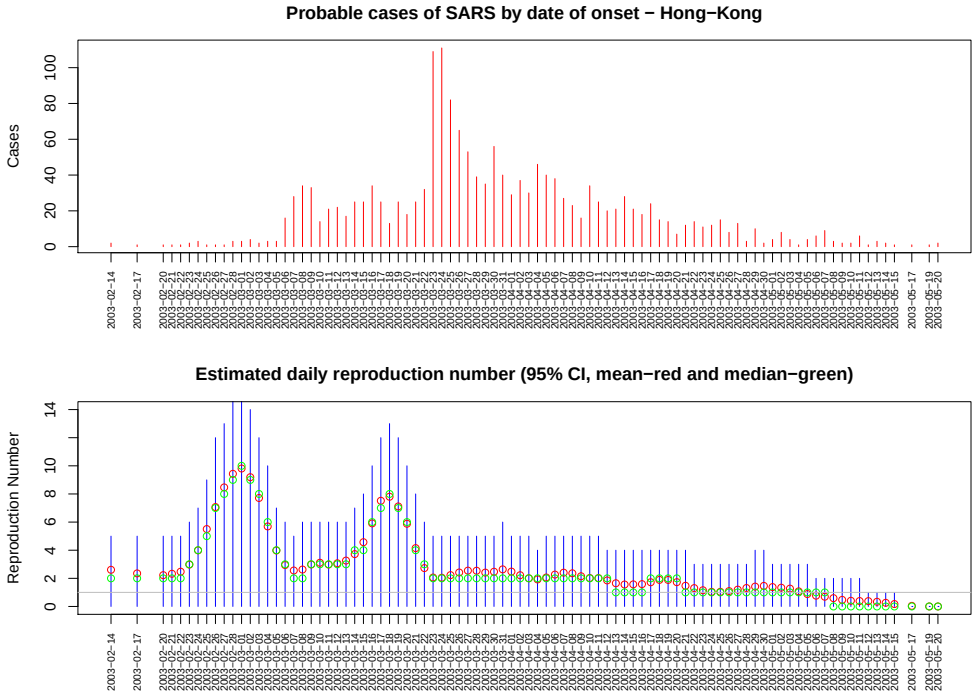

Figure 7: Top panel: Probable number of SARS cases in Hong-Kong by onset date for the 2-14-2002 to 5-20-2003 time period. Lower panel: The W-T estimate for the daily reproduction number. Red/green circles are the mean/median values and the blue bars mark the 95% CI. The horizontal grey line is drawn at $R = 1$, the critical value for the epidemic spread.