The Global Epidemic Simulator

Release 1.0 BETA

Wes Hinsley, Pavlo Minayev

Stephen Emmott, Neil Ferguson

MRC Centre for Outbreak Analysis and Modelling
A Microsoft Research Project

1. Synopsis

The Global Epidemic Simulator is a platform for modelling directly transmissible human diseases such as influenza, and testing intervention policies that may reduce the duration or severity of an outbreak of such a disease. As in previous work[i], our model is spatial and individual-based, using various geographical and demographic data sources to place households at absolute locations, populated with individuals of specific ages who each attend schools or workplaces situated in their vicinity. The significant challenge of this project over previous work is the global nature of simulation. Memory and performance requirements motivate us towards a distributed-computing approach, while the availability of reliable data for every country is also challenging.
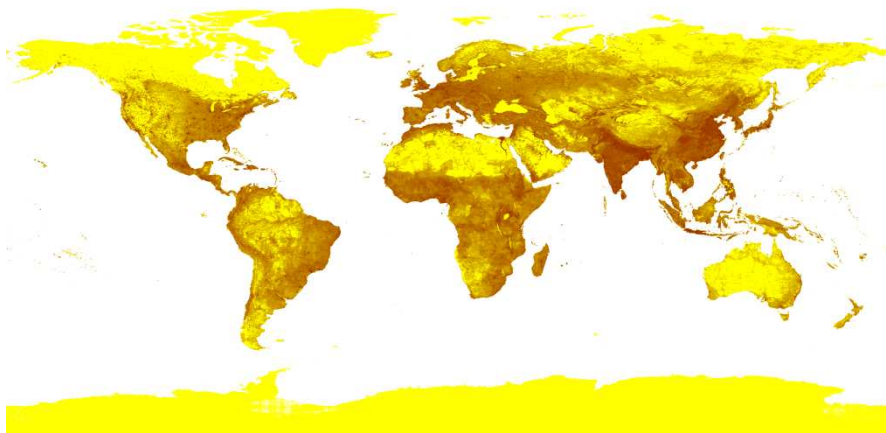
The Synthetic Population

Figure 1: Landscan 2007 dataset, population density plotted on a logarithmic scale. (Lighter=lower density)

We used the Landscan[ii] dataset which provides population counts at a resolution of 30 arcseconds longitude and latitude – around 1 square kilometre near the equator. This is combined with demographic data specifying the number of people in a household, and their age distributions; this information is harder to find, and is accumulated in a somewhat ad hoc way from the international and governmental statistics sites (United Nations Statistics Division[iii], EUROSTAT[iv], DHS (USAID)[v], UK Office for National Statistics[vi], US Census Bureau[vii], etc.). Individuals have incubation periods, infectious periods and infectiousness profiles that can be fixed, or defined with statistical distributions.

Households, workplaces and schools are then grouped by administrative unit (NUTS level[viii]); these are geographical boundaries at varying resolution, freely available from GADM[ix]. Data for some countries is only down to county-level boundaries; others such as Kenya have detail to NUTS level 5. Any multi-level combination of units can be set up. The parents of child-units are automatically included, as is a catch-all global administrative unit, hence a hierarchy is in place for parameter overriding, and output stratified by NUTS level. Interventions can applied, and output can be logged of these units at any level.
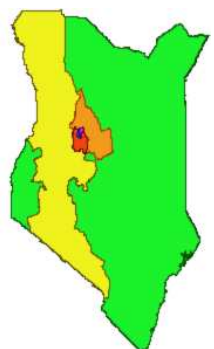

Figure 2: A rare example of a level 5 admin unit hierarchy in the GADM database:- Angata Nanyukie, Lorroki, Samburu, Rift Valley, Kenya.

Transmission methods

An infected individual can contact and possibly infect a susceptible individual either through (1) their household, (2) their workplace or school, or (3) through random contacts in their local vicinity, or on journeys they make, including long range travel. Transmission coefficients for each of these methods can be defined spatially.

School, workplace and household transmission is treated with fairly simple probabilistic rates depending on susceptibility and infectiousness; community contacts are more complex. To avoid having to calculate a prohibitively large number of individual-to-individual contact probabilities, we use a quantised grid of patches, and calculate patch-to-patch probabilities using a gravity model. Each node contains a set of non-overlapping local patches and remote patches: local patches contain households and people, and are always at resolution 20x20 Landscan cells, whereas remote patches contain only a population count, and a reference to another node which will contain the households and people for that region.
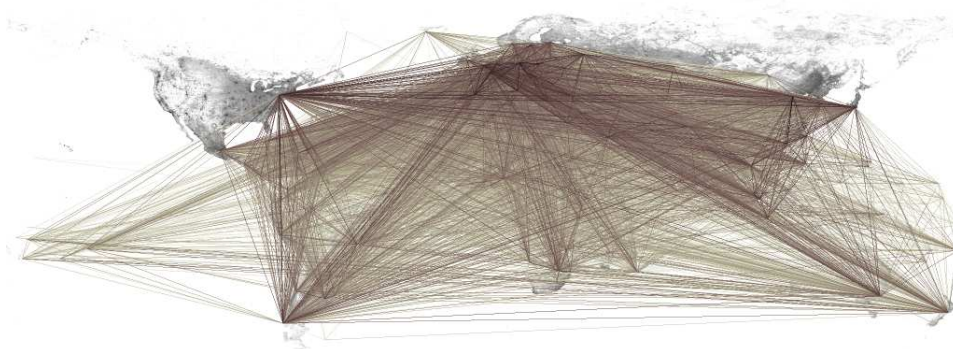
Long-range travel

Figure 3: Border crossing links in the WTO 2004 dataset, plotted on a logarithmic scale.
(Darker = more crossings). WTO is country-to-country: points used above are capital cities.

The World Tourist Organisation Factbook[x] provides annual country-to-country counts of people crossing borders. While a higher temporal and spatial resolution (and general completeness) would be desirable, it has proven difficult or prohibitively costly to acquire such data. Never-the-less, the WTO data, combined with country population can be used to approximate crude probabilities of a randomly picked person being a traveller and if so, to pick their origin or destination. Travellers then use the community contact algorithm, possibly at home and abroad, so infected people can potentially spread their infection overseas, and susceptible travellers can catch infections from overseas and infect people at home.

Interventions

A range of intervention policies can be added, which turn on and off at specified times, or when specified thresholds of cases are crossed. These include restricting country borders, applying blanket travel restrictions, antivirals and prophylaxis, treatment of infected people, isolation and quarantine, closure of schools and workplaces, and vaccination. These have various effects on the susceptibility and infectiousness of individuals, and are applied with varying success or compliance rates.

High Performance Computing Issues

OpenMP is extensively used throughout the main loops and much of the initialisation. Various thread-indexed arrays and stride-based iteration methods are used to eliminate the need for mutual exclusion zones as much as possible.

Distributed Computing Issues

We use MPI to communicate between nodes where necessary. Three consensus-like algorithm issues arise. Infected hosts can try to make community contacts across node boundaries. An MPI request message is sent and the success of establishing those contacts is not known until the following time step when a reply message is received. The infector therefore provisionally schedules enough local contacts such that if every remote contact were to fail, there will still be sufficient local contacts. Successful contacts and their ordering are then broadcast and the unnecessary ones cancelled.

The second issue arises where a single workplace or school may have individuals that live in households on multiple nodes. The establishment is then duplicated on all relevant nodes, while the individuals are not duplicated. MPI messages are used to synchronise infection events across the duplicated establishment copies, and to apply any interventions triggered.

Similarly, the third issue involves administrative units that may span across node boundaries; in this case, infection information must be synchronised so that policies applied to the unit are activated on all applicable nodes at the same time.

Output

The simulation can write results to a database table via ODBC, or to flat files. In each case, separate streams of output come from each node, at administrative unit resolution, for any selection of units. The simulator can also produce PNG files for assembling into a movie for quick visualisation, rather than detailed analysis.
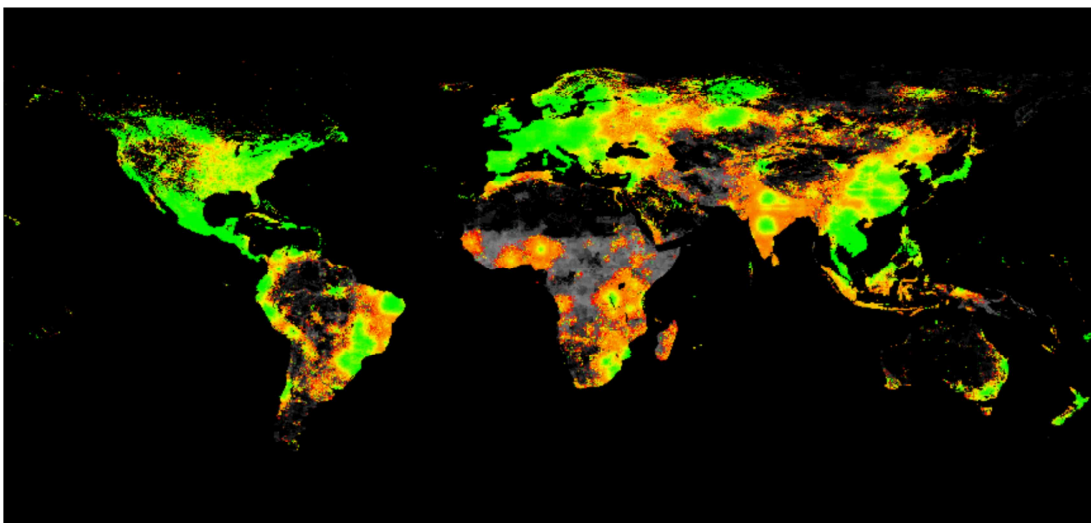

Figure 4: Snapshot during an epidemic seeded in Vietnam.
Green indicates recovered hosts, red indicates infected hosts.

2. Software Pre-requesities

Required:

- For a global run, a MPI-capable cluster of machines. The exact memory requirement varies depending on the job. This release comes with configurations we have run on 3 nodes (max usage 300Gb RAM per node), and 9 nodes (max usage 128Gb per node). Contact us to discuss different configurations.

- A 64-bit platform is required for job execution and data preparation; it can be Windows or Linux based.

- Storage space for population dataset: 140Gb, with a further 140Gb per customisation.

- For job building and customisation, a Java Runtime Environment 7.0 or later is required.

- For job launching, the default launch scripts are for Microsoft HPC 2008 for Windows, and PBS for linux. These are easily modified. A model can be developed on Windows, but launched on a linux cluster, and vice versa.

- A download of either the Landscan dataset, which has license restrictions, or the freely available GRUMP dataset. Other data files will be downloaded during setup.

Optional:

- For recompiling the simulation code and other tools, a C++ compiler such as GCC, Microsoft Visual Studio, or the Intel C++ compiler. OpenMP support is highly recommended. See notes on compilation. MPI libraries will be required, depending on cluster platform.

- For recompiling the customisation tools, a Java 7 compiler, such as Oracle's JDK is required

- An ODBC-compliant SQL server can optionally be used for storing results of runs.

3. Preparing the Platform.

Throughout this section, instructions from a windows command prompt will be written in red, and equivalents for a linux bash shell will be written in blue. It is assumed at this stage that the pre-compiled binaries will be used. See the later section should you need to recompile. On starting each section, it is assumed the current directory is the root of the release.

- Download and decompress the ZIP file of the release. If you are running on Linux, then prepare the permissions of the scripts and executables with:-

```
chmod 755 fixPermissions.sh
./fixPermissions.sh
```

- Provide the population raster files in data/RequiredData. These files are ESRI grids, which come as a pair of files: one header (HDR) file, and one flat binary (FLT) file. We cannot provide these downloads, but they can be easily obtained.

  The GrumpV1 dataset is available at http://sedac.ciesin.columbia.edu/gpw/global.jsp - ensure you download the 30" resolution version. You will need to register, but it is free to do so. Choose the ".bil" file format, and open with an unzipping tool (such as www.7-zip.org). The "readme.txt" will help you choose whether you want the "adjusted" or "unadjusted" version – extract the appropriate .HDR and .BIL file, then rename the .HDR to **population.hdr**, and the .BIL to **population.flt**.

  An alternative which we have used for development is Landscan, available at http://www.ornl.gov/sci/landscan/ . This however requires a license agreement with ORNL. The download then comprises a binary grid, with **.hdr** and **.flt** files to extract and rename to **population.hdr** and **population.flt**.

- Run the first data preparation script:-

```
cd bin-w64/GetAdminUnits          cd bin-linux/GetAdminUnits
get_admin_units                   ./get_admin_units.sh
```

  This will launch a Java executable which downloads the entire set (1Gb) of administrative units from the GADM/DIVA-GIS site into data/AdminUnits. It will then reduce and compress them into a single 52Mb zip file, and create various small accompanying metafiles. Finally, it will create a 3.7Gb raster country file in the ESRI file format (FLT flat file and HDR header file), in data/AdminUnits. This is used for defining borders for the synthetic population generator.

  Note that if you re-download the admin units at a later date, it is possible you will need to edit your models, since the admin unit ids may change.

- Generate the synthetic population.

  This is a lengthy and memory intensive job, which will usually be best to run on a cluster. Our scripts are customised towards this.

  ```
  cd bin-w64\SynthPopul          cd bin-linux/SynthPopul
  ```

  Now examine the file run_single.bat or run_single.sh in a text editor such as EditPlus on Windows, or pico/nano on Linux. Change the definition of **REPO** to define the path to the release directory, which must be network accessible. Set the definition of **CLUSTER_CMD** so that on your particular cluster, adding a network path to a batch file will cause that batch file to be submitted for execution to the cluster. For Windows HPC, this may be job submit /scheduler:headnode /jobtempate:template /numNodes:1 or for PBS on Linux, it may be qsub –qQueue. In the unusual case that you want to try and run the synthetic population generator on your local machine, then this can be left blank. The process will require around 40Gb of RAM for large countries.

  Now run the script:-

  ```
  generate_synth_pop.bat          ./generate_synth_pop.sh
  ```

  This will call the run_single script once for each country in the global simulator's domain. Each call will add files to the directory data/SynthPopul. You now need to wait for the synthetic population to be fully created before proceeding to the next step.

- Generate a raster-grid from the new synthetic population.

  The synthetic population is a distribution of household and establishment files, rather than individuals. We now need to create a rasterised version of the households, which we will then use to create the node distributions.

  ```
  cd bin-w64\CombineSynthPopul    cd bin-linux/CombineSynthPopul
  combineSP                       ./combineSP.sh
  ```

  This creates a new binary grid, population.bin in data/NodeConfigs.

- Create the node distributions.

  This will generate the initialisation files for our default 3x256Gb node configuration script. Tools for creating new node distributions will be added in a later release; contact us if you need a different configuration in the meantime.

  ```
  cd bin-w64\PatchFileMaker       cd bin-linux/CombineSynthPopul
  makePatches                     ./makePatches.sh
  ```
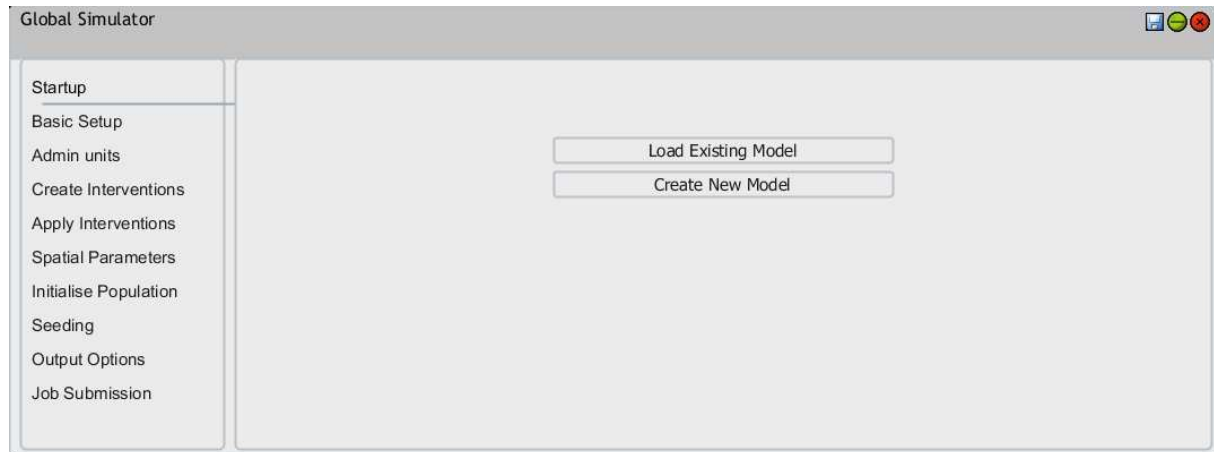
  If all of these steps were successful, you are now ready to run the Job Creator tool.

4. Using the Job Creator

- Run the job creator interface with:-

```
cd bin-w64\JobCreator          cd bin-linux/JobCreator
gsimkit.jar                    ./gsimkit.jar
```


Above is the starting screen (with colours adjusted for better printing). On all pages, the three icons on the top right allow you to save any changes, minimise the window, or close the job creator.

As you create a model, you will visit each page listed on the left. The first page is for picking a model to customise, or creating a new one.

- Create or Load a model
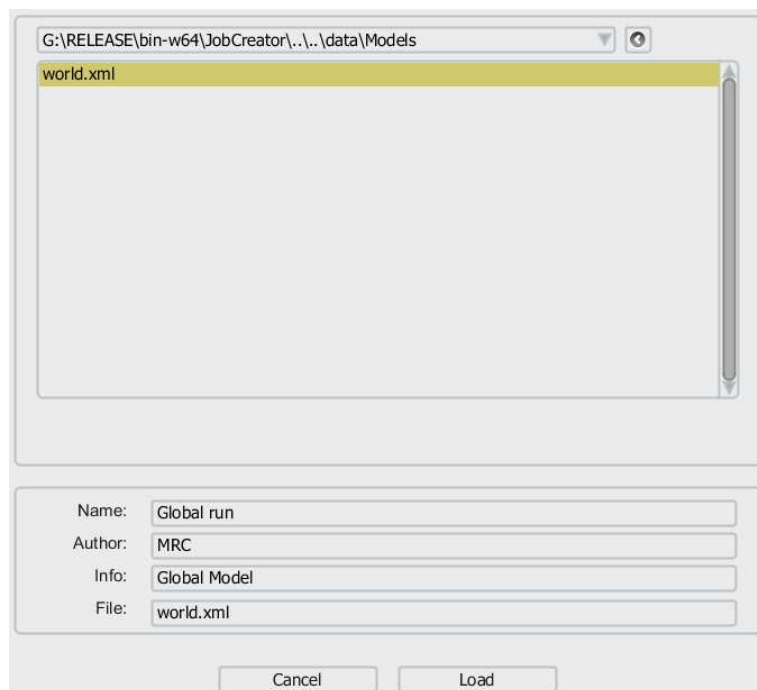
Clicking on load shows the available models in the default folder.

Clicking on a model shows the meta-data at the bottom.

Click on to move up a directory level; directories will also be shown in the window.

Click on Load to load the model, or Cancel to return to the start.

The window for creating a new model is similar, except you can edit the information about the model, including the file name.

Clicking on an existing model in the lower window and changing the filename will make a new copy of that model for editing.

Click create to continue editing the model, or cancel to return to the start.

- Basic Settings: Latent period, infectious period and infectiousness

On this page, functions for the latent period, the infectious period and infectiousness over time can be set. Either infectiousness or the infectious period (or both) must be fixed; they cannot both be distributed. Tick the appropriate box to choose which is fixed.

You can then click on each of the functions to define. The distribution functions available are Fixed constants or parameterisable Gamma, LogNormal or Weibull, with offsets and tail truncation. The graph preview window can show the probability distribution function, the cumulative distribution function or inverse cumulative distribution function.

- Administrative Units
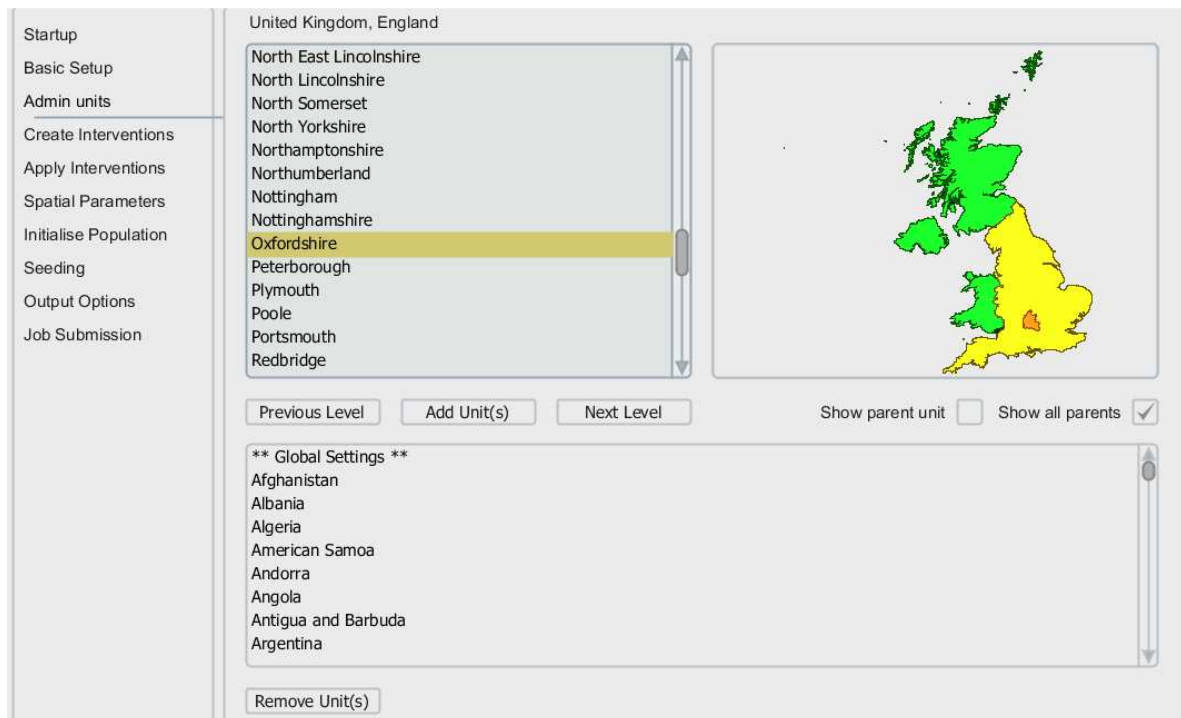
An administrative unit serves the following purposes:-

- All geographically varying parameters are specified per-admin-unit.
- Interventions are turned on/off at admin unit level
- Interventions are triggered by admin-unit-level statistics.
- Output is logged per-admin-unit.

The maximum resolution of administrative unit varies from country to country. Some of the smaller islands may have only level 1, whereas Kenya is currently unusual in having level 5 detail.

Every model will have a "Global" catch-all unit with default settings and global logging; all individuals, households and places in the simulation that do not belong to a more detailed administrative unit will inherit these global defaults for their parameters. Hence, the simplest model you could run would have just the global administrative unit, and all geographic regions would then use the same parameters and interventions. You would only be able to log the global bulk statistics for simulations generated with that model.

Choose the units you would like in the top window. They will be previewed in the graph window. You can select multiple units by holding shift or control while clicking. The buttons allow you to go to the next (more detailed) level where available, or the previous one.

The bottom list shows your current selections. If you select some level 3 unit, the job builder will automatically select the parents of that unit, and you can only remove a unit if there are no children of that unit in your selection list.

- Creating Interventions

In this page, generic interventions are created. They are applied to different places in the following page. The platform currently allows country border controls, treatment of symptomatic patients, prophylaxis of uninfected individuals, mass vaccination, quarantine, closure of schools and workplaces and blanket travel restrictions.

All are parameterised in various ways – see the technical model section later for explanations of the interventions. Each intervention is turned on, and turned off in particular ways that can either be at a fixed day number, or when the number of clinical cases reaches some threshold number of cases(or some percentage), in either a daily period, or a cumulative total.

The top section shows the properties of the intervention being edited, and the list at the bottom shows the interventions currently existing in the model. You can adapt existing interventions by clicking on the name in the bottom list and choosing Edit, then changing the name at the top.

- Applying interventions

The next page allows interventions to be applied to any admin unit by ticking the boxes. Siblings (units of the same level with the same parent), can be grouped together, so clicking one sibling will activate (or deactivate) that intervention policy for each of them.

- Spatial Parameters

Here, parameters that vary spatially can be specified for each admin unit. A table describing the parameters is given later. One set of parameters can be applied to a range of admin units at once by selecting them on the right side, either by clicking with shift or control, or by using the "Select Siblings" or "Select All" buttons.

- Initialise susceptibility

Initial susceptibility can be defined for arbitrary age groups in this table. Add new age bands by filling in the range boxes and clicking on "Add".

To remove age bands, merge them together by clicking on the first band, hold shift and click on the last band. Then click the "Merge Groups" button.

Ticking the "Link all ages" button will apply any susceptibility change you make to every age band.

- Seeding

The initial seeding of the simulation is done next. Two seeds for the random number generator must be specified, along with the location, time, and number of infections that start the epidemic. The list on the right contains most major cities; click on them populates the longitude and latitude boxes. The day number of the simulation allows seeding to be distributed over time if required. Then specify the number of cases to seed at that location. The '+' button adds this entry to the list of seed, and the 'X' button removed existing entries.

- Output Options

Output can be written to an SQL database (if configured), to flat tab-separated files, and as bitmap (PNG) images for appending into a movie. You must choose which administrative units you would like statistics for on the left; you may find the "Link all" and "Link siblings" tickboxes helpful for selecting a range of admin units in one go. On the right, select which forms of output you require.

For the database option, you must have an ODBC connection set up in your operating system, with the server name given here, along with the database table for results. The flat files and PNG images must be given an output path that is network accessible, since runs will most often be run on a remote cluster.

- Job Submission

On the final page, the settings for job submission need to be given. The job-type is the node configuration required for this job. The supplied configuration is for a global simulation spread across 3 cluster nodes. Other configurations may be added in future releases. The platform is the target on which the final simulation will run, which can be Windows or Linux.

Submission then happens in two stages. **Firstly**, the synthetic population files you created when setting up the platform need to be copied and modified for the set of administrative units you have chosen for this model. **If you plan to do multiple simulations using the same set of admin units, then you will only have to do this process once**, and you can re-use the modified synthetic population again.

You then need to specify the path to the original synthetic population files, a path where the modified files should be places, a working directory where scripts and executables will be put, and finally the command used locally to launch jobs on a cluster, since the modification process is memory intensive. You can use the "…" buttons to select a path using a file browser.

Next, click the "Generate map file", which generates a raster version of the administrative unit polygons in the modified SP directory. It may take a few minutes to generate this. After this, click the "Prepare scripts" button, which creates launch.bat and run.bat (or .sh equivalents) in the working directory. Finally, the "Launch Modifier" button calls the launch script.

The **second** and final submission stage, is running the simulation itself. You need to supply a path to the modified synthetic population, a working directory, and a launch command as before. Then the second "Prepare scripts" button writes the launch and run scripts, copies the necessary executable files and dependencies, and writes the parameter file. Finally "Launch Sim" will call the launch script.

5. The parameters

The following parameters can be specified for each admin unit in the simulation.

Name	Description
P_symptomatic	Probability of an infection being symptomatic
P_sympt_detect	Probability of a symptomatic being clinically detected
Mul_sympt_inf	Multiply transmission coefficients if infection is symptomatic
P_severe	Probability of a symptomatic infection being severe
P_severe_detect	Probability of a severe infection being clinically detected
Mul_severe_inf	Multiply transmission coefficients if infection is severe
B_spatial	Transmission coefficient for community contacts
Spatial_krn_a	Spatial Kernel function $f(d) = \dfrac{1}{1+\left(\frac{d}{a}\right)^b}$ if d<cut, otherwise 0.
Spatial_krn_b	
Spatial_krn_cut	
B_household	Transmission coefficient for within households
B_Pn	For each place type n, {0,1,2,3}, transmission coefficient.
P_Pn_group	Within-group transmission probability for each place type n.
Pn_sympt_absent	Probability that symptomatic individual is absent from place type n.
Pn_sympt_absent_com	For symptomatic absentee, multiply community contact transmission.
Pn_severe_absent	Probability that severely infected individual is absent from place type n.
Pn_severe_absent_com	For severely infected absentee, multiply community contact transmission
Seasonal_lat_max	$L = \dfrac{max+min}{2} - \dfrac{max-min}{2}\cos\left(\dfrac{2\pi latitude}{180}\right)$
Seasonal_lat_min	Temporal effect= $1.0 + L(\cos(offset + \omega T\_days))$
Seasonal_temporal_offset	(multiply transmission coefficients by result)

6. The interventions and their parameters

Interventions can be applied flexibly at admin unit level. Below are their configurable parameters.

Border Controls	
P_deny_entry	Probability that entering country will be refused
P_deny_exit	Probability that leaving country will be refused
Treatment of Symptomatic Individuals	
Mul_inf_sympt	Multiply infectiousness of symptomatic by this factor
Trt_delay	Delay from symptoms to treatment (days)
Trt_duration	Treatment Duration (days)
Prophylaxis of Susceptible Individuals	
Mul_susc_proph	Multiply susceptibility of uninfected by this factor
Mul_inf_proph	Multiply infectiousness of post-prophylaxis infection
Mul_sympt_proph	Multiply probability of post-prophylaxis symptomatic infection
Proph_delay	Delay from symptoms to prophylaxis (days)
Proph_duration	Duration of prophylaxis (days)
Proph_hh_coverage	Proportion of households prophylaxed
Proph_pl_coverage	Proportion of class/workgroup prophylaxed
Vaccination	
Mul_susc_vacc	Multiply susceptibility of uninfected individuals by this factor
Mul_inf_vacc	If still infected, multiply infectiousness by this factor
Mul_sympt_vacc	If still infected, multiply symptomatic probability by this factor
Vacc_delay	Delay between vaccination and protection (days)
Vacc_covergae	Proportion of unit vaccinated
Case Quarantine	
Q_mul_B_con	Multiply school/workplace transmission by this factor
Q_mul_B_hh	Multiply household transmission by this factor
Q_period	Period of quarantine (days)
Q_delay	Startup delay (days)
Q_compliance	Probability of compliance with quarantine

Place Closure	
Pc_threshold	Threshold cases for closure (absolute number, or %)
Pc_delay	Delay to implement closure (days)
Pc_period	Closure period (days)
Pc_hh_mul	Multiply member's household transmission by this factor
Pc_cc_mul	Multiply member's community (spatial) transmission by this factor
Blanket Travel Restriction	
Bt_max	Maximum travel distance (km)
Bt_enforce	Probability of enforcing restriction.

Interventions can switch on and off at either fixed times, or when defined thresholds are crossed. The thresholds are numbers of clinical cases (or percentages within the unit), encountered either daily, in the last ten days, or cumulatively summed.

7. The results

Flat files produced by the simulation are tab-separated text files, from each node, named "out_n.txt", for each node n. Database tables contain the same fields; column names are in brackets below.

1. Time (hours into simulation)   [time_hrs]
2. Admin Unit ID.   [Unit_id]
3. People making contacts   [Contact_makers]
4. New community cases   [Comm_cases]
5. New household cases   [HH_cases]
6. New nursery school cases   [P0_cases]
7. New primary school cases   [P1_cases]
8. New secondary school cases   [P2_cases]
9. New workplace cases   [P3_cases]
10. New community infections   [Comm_infs]
11. New household infections   [HH_infs]
12. New nursery school infections   [P0_infs]
13. New primary school infections   [P1_infs]
14. New secondary school infections   [P2_infs]
15. New workplace infections   [P3_infs]
16. Current symptomatic infections   [Current_Sympt]
17. Current non-symptomatic infections   [Current_NonSympt]

Here we define "cases" to be clinically detected cases, and "infections" to be the sum of all detected and non-detected, symptomatic and non-symptomatic infections.

A further text file generated at launch time, admin_lookup.txt, is a tab-separated file giving the numerical ID, and the string name for each administrative unit.

8. Compilation

Scripts are provided for recompilation in the form of Windows batch files, and Linux shell scripts. In either case, the presence of GCC (g++) and javac are assumed. GCC (g++) must be recent enough to support the –fopenmp switch, and javac must be equivalent to OpenJDK 7.0 or later. The provided libraries and include directories should be sufficient to ensure compilation on all linux and windows platforms.

To recompile all of the source code, and replace the executables in the bin-w64 or bin-linux folders with new copies:-

```
cd src                    cd src
combine_all               ./combine_all.sh
```

Note that:- we believe the Intel C++ compiler to produce the highest performance code at present. The binaries we have provided in the release were produced from this compiler, and have the necessary redistributable DLL files present.

If you recompile using a different compiler on Windows, you will have to ensure that the appropriate DLL files are available. Usually the simplest way is to copy them into the same folder as the executable.

9. The Simulation Model

(i). Community Contact Model

The number of community contacts made by an infected individual in their infectious period is given by:-

$n_{contacts}$ ~ poisson (T_{inf} * $B_{spatial}$)
* $c_{community}$ if school/workplace is closed
* $q_{community}$ if household is under quarantine
* severity_multiplier   if the case is classed as severe.

$B_{spatial}$ is the transmission co-efficient (per day), and T_{inf} (days) is the infectious period of the infected individual. The number of contacts is then sampled from a poisson distribution. Note that $B_{spatial}$ may have a different value for each user-configured administrative unit.

A rejection algorithm is used to pick the contacts to be made. Contacts are assumed to be made uniformly throughout the infectious period. For each contact attempt:- if the infector is called i, and the susceptible j:-

$patch_j = q_{ij}$ (pick random patch from pre-computed distribution)

Note that if $patch_i$ and $patch_j$ are in different countries, then the attempt to contact will fail at this point with user-defined probability p_deny_exit, if a border control intervention is applied.

D_{ij} = shorted distance between $patch_i$ and $patch_j$
r_{ij} = absolute distance between i and j

If blanket travel restrictions are applied, and r_{ij} exceeds the threshold distance $bt_{threshold}$, then the attempt to contact will fail at this point with user-defined probability p_{deny_bt}.

$S_{ij} = F(r_{ij})/F(D_{ij})$
If rnd()<S_{ij} accept contact, else reject contact

Where F is a kernel function of the form F(d)=1/(1+(d/a)^b). Parameters a and b are user-defined for each administrative unit, along with a maximum distance cut-off parameter. If the contact is accepted, then the new contact becomes infected with probability

P_{infect} = susceptibility(j) * infectiousness(i,t)

Where susceptibility and infectiousness are defined below.

(ii). Global Travel

Global travel is implemented as an extension of the community contact algorithm. When a new contact becomes infectious, their location throughout their infectious period is decided. The WTO dataset (see later), allows us to crudely estimate two probabilities:-

$p_{traveller}$: a given person in a given country will travel to another country.
$p_{visitor}$: a given person in a given country will be a visitor from another country.

On probability $p_{traveller}$, or $p_{visitor}$, we create a travel plan for the newly infected individuals, which involves the infected host being away in a specified country for all or part of their infectious period, while ensuring they are present with the infector, at the time of the infection. The time spent away, T_{away}, is used to reduce household and school/workplace contact rates proportionately.

This is not done, however, if a border control intervention is active, in which case this is ignored with probability bc_{deny_exit} or bc_{deny_entry}, depending on whether we are making a traveller's plan or a visitor's plan respectively.

We also assume that the blanket travel restrictions intervention prohibits global travel, if the associated probabilities for that intervention are met.

(iii). Household contacts

In each timestep, establishing household contacts is modelled with a Bernoulli trial for each member of the household other than the infector. The probability of the infector i contacting a household member j is:-

$p_{contact}$ $= 1-exp((-B_{hh}*(T_{inf} -T_{away}))/n_{members}-1)$
         $* c_{household}$ if workplaces/schools are closed by intervention
         $* q_{household}$ if household is under quarantine.
         $*$ severity_multiplier if the case is classed as severe.

Where B_{hh} (per day) is the household transmission coefficient, T_{inf} (days) is the infectious period of the infector, and $n_{members}$ is the number of members of the household. Bhh can be given different values for each administrative unit.

Given that contact has occurred, the probability of member j becoming infected is:-

P_{infect} = susceptibility(j) * infectiousness(i,t)

where susceptibility and infectiousness are defined later.

(iv). School/workplace contacts

In each timestep, establishing household contacts is modelled by sampling a Binomial distribution to calculate how many members of the school/workplace should be contacted. The probability of contact for workplaces is **zero** if the school/workplace is closed, otherwise:-

$p_{contact}$ $= 1-\exp((-B_{workplace}*(T_{inf}-T_{away}))/n_{members}-1)$
$* q_{social}$ if the infector's household is under quarantine.
$* severity\_multiplier$ if the case is classed as severe.

$n_{contacts}$ $\sim binomial(p_{contact}, n_{members}-1)$

The members of the establishment to be contacted are chosen at random. For each, a successful infection takes place with the probability:-

P_{infect} $= susceptibility(j)*infectiousness(i,t)$

(v). Susceptibility

Initial susceptibility is set for each of a user-defined number of age bands, and is assumed to apply to people at all locations. Instantaneous susceptibility is then:-

Initial susceptibility $* prophylaxis_{susc} * vacc_{susc}$

Where prophylaxis may be occurring in the susceptible's household or workplace, and vaccination may have previously occurred.

(vi). Infectiousness

Infected hosts have an infectiousness profile varying over time, modelled in a user-defined way – a gamma probability distribution function for example. The infectiousness at a given time is:-

$Inf\_profile(t) * treat_{inf} * prophylaxis_{inf} * vacc_{inf}$
$* relative\_inf\_symptomatic\_multiplier$

Where the multipliers for the first three are only included if relevant intervention policies are active in the infector's administrative unit, and the last is only included for symptomatic infection.

(vii). Other factors regarding infection

When a susceptible becomes infected:-

Their infection is symptomatic with probability

$$p_{symptomatic} \quad * \quad vacc_{sympt} * prophylaxis_{sympt}$$

where vaccination and prophylaxis have user-defined modifiers of the probability of being symptomatic if infection still occurs.

Their infection is clinically detected with probability $p_{case\_detection}$

Their infection is severe with probability p_{severe}

9. Authors and Acknowledgements.

The Global Epidemic Simulator was developed by Dr. Wes Hinsley and Dr. Pavlo Minayev, supervised by Professor Neil Ferguson, at the MRC Centre for Outbreak Analysis and Modelling, Imperial College London, and Professor Stephen Emmott, Microsoft Research Cambridge.

The project was funded by Microsoft Research, Cambridge.

The border crossing data for global travel data were taken from the World Tourism Factbook 2004, used by kind permission of the World Tourism Organisation. © UNWTO, 9284404111. http://www.unwto.org.

The code for producing PNG files is LodePNG, a free open source library by Lode Vandevenne. See original comments in lodepng.cpp.

The Java Job Creator uses the Jaxen/Dom4J free libraries for XML support. See http://jaxen.codehaus.org/ and http://dom4j.sourceforge.net, and the Java Statistical Classes (http://www.jsc.nildram.co.uk/) for generating the distributions.

10. References

[i] Strategies for containing an emerging influenza pandemic in Southeast Asia. Ferguson et al. *Nature* **437**, 209-214 (8 September 2005) | doi: 10.1038/nature04017

[ii] Landscan is available here: http://www.ornl.gov/sci/landscan/, however, licensing issues apply. An alternative data source is the GRUMP alpha project: http://sedac.ciesin.columbia.edu/gpw/global.jsp, providing comparable data from 2000 at the same resolution as Landscan.

[iii] United Nations Statistics division: http://unstats.un.org/unsd/demographic/products/dyb/dyb2.htm

[iv] EUROSTAT: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[v] DHS (USAID): http://www.measuredhs.com/

[vi] UK Office for national statistics: http://www.statistics.gov.uk/hub/population/index.html

[vii] US Census Bureau: http://www.census.gov/

[viii] NUTS level: http://en.wikipedia.org/wiki/NUTS

[ix] Global Administrative Area: http://www.gadm.org , partly duplicated at http://www.diva-gis.org

[x] The WTO Factbook: http://www.e-unwto.org/content/v486k6