SurfaceBuilder247: User Guide

Purpose

SurfaceBuilder247 is software designed to facilitate the creation of gridded population distribution models for specific times and dates. It is a very substantial development of the algorithm used in the *SurfaceBuilder* software originally written by David Martin and available at http://www.public.geog.soton.ac.uk/users/martindj/davehome/software.htm. SurfaceBuilder247 processes a series of input centroid locations, such as census output area (OA) centroids, schools, hospitals, etc. with associated population counts and time profiles and models these onto a regular geographical grid at a specified target time and date. A background layer allows weighting of locations according to their probability of containing population and may be used to represent the transportation network and areas to be excluded from the modelling such as the sea. The program proceeds by redistributing the total population represented by a series of "origin" centroids onto the grid, taking account of relevant centroid locations and the background layer, so as best to reflect the demand exercised by each location at the specified time. Unlike conventional population mapping tools, including the original SurfaceBuilder, which assume all population to be present only at residential locations, which in effect is a "night time" population representation, this approach allows the modelling of any desired night time or daytime population distribution.

SurfaceBuilder 247 is a principal output from the ESRC-funded "Population 24/7: space-time specific population surface modelling" project (Award Number RES-062-23-1811). Further information about the project (including relevant papers and presentations) can be obtained from the ESRC website at http://www.esrc.ac.uk/my-esrc/grants/RES-062-23-1811/read and the project website at http://www.southampton.ac.uk/geography/research/projects/space_time.page. The award holders were David Martin, Samantha Cockings and Samuel Leung. The SurfaceBuilder247 programme specification was originally written by David Martin and programmed by Jason Sadler and Hugh Darrah. Use of SurfaceBuilder247 requires preparation of an extensive input data library, but a selection of exemplar output models for England and Wales are available using the *Population 24/7 Download* application and this is the recommended starting point for a new user. As the exemplar models incorporate data sourced from the ESRC Census Programme http://census.ac.uk the download application requires users to be registered with the census programme for use of 2001 census aggregate statistics and postcode directory datasets. No registration is required in order to use the SurfaceBuilder247 tool itself, but the user must ensure that they have appropriate access rights to use all the input datasets which may be required for their intended modelling.

User guide

Installation

SurfaceBuilder247 has been written in Visual Basic .NET and is supplied as a zipped folder containing an executable file (STSurfaceBuilderAapp.exe) and two library files (Microsoft.Office.Interop.Excel.dll and Microsoft.Vbe.Interop.dll) required by the program. The software does not require an installation or setup process to be run, but the program files should be unzipped as part of a file structure containing a data library as described below. The software has been tested by the project team on PCs running Windows XP, Vista and Windows 7. As the application is computationally intensive, processor speed and available memory have a significant impact on modelling performance for large datasets.

Folder structure

A folder structure should be set up as follows in order that all the various files required by the program can be correctly located. The entire system should be contained in a folder on the C: drive of the PC called C:\UC1264_Pop247 Inside this folder, four subfolders are required:

- *Docs* is used to hold relevant documentation such as this User Guide, file format references, etc.
- *Dates* may be required to hold date reference information.
- *Program* contains the program files noted above, which should be unzipped to this location.
- *Data* is used to build the data library on which the program is based.

The Data folder requires further subfolders which contain the different file types either required or generated by the program as follows:

- *BckGrnds* contains pre-prepared background map layers
- *DataLogs* contains data log files generated by SurfaceBuilder247
- *Dests* contains destination centroid data files
- *Origins* contains origin centroid data files
- *Results* contains results files
- *RunLogs* contains run logs generated by SurfaceBuilder247
- *SessionParas* contains files recording session parameters for SurfaceBuilder247 runs. These may be saved from the program itself or generated by the user.
- *TimeSeries* contains the timeseries library files
- *ValidationLogs* contains file validation logs generated by SurfaceBuilder247

Preparing a data library

Preparing for a SurfaceBuilder247 run requires the user to assemble an extensive library of data files in the Bckgrnds, Dests, Origins and TimeSeries folders. This is the most time-consuming part of the work and requires extensive planning and a full understanding of the spatio-temporal modelling concepts.

The background map layers are ascii (.txt) files set out in the asciigrid format used by ESRI's ArcGIS software, at the same cell resolution as the desired output grids and covering the entire area to be modelled. These grids contain weighting values to indicate the relative likelihood of population in the transportation system being located in each cell and nodata values to indicate cells in which it is no population should be located. Such a layer may typically be created in a GIS by rasterising a transportation network with associated capacity values and overlaying a layer representing areas of sea, open water or other areas from which modelled population must be excluded. The file contains a header in the form:

```
ncols           3500
nrows           3500
xllcorner       0
yllcorner       0
cellsize        200
NODATA_value -9999
(Followed by individual cell values in row primary order...)
```

The main inputs to SurfaceBuilder247 modelling are the files representing population destinations and origins. These essentially represent all locations which may be considered "containers" of population at some time within the reference frame of the modelling. Examples of origins are census output area (OA) centroids, whose total population counts sum to equal the total population of the model. Examples of destinations are any locations which have no resident population but at some time may contain population such as workplaces, educational establishments, health care facilities, etc. The levels of detail of origin and destination centroids may be expanded as far as available data and modelling requirements allow but it is important that within a single data library the entire population is only accounted for once within the origin centroid set and that destinations similarly avoid the possibility of double counting. For example, origin populations may be referenced to census OAs or to postcode locations, but only one of these referencing systems should be used within the same geographic extent, such that the entire population is accounted for once and once only. Similarly, a destination file may contain workplaces referenced by OA centroids, or may contain every known workplace separately identified by exact grid coordinates, but it is essential that within a single data library each workforce is only represented once, hence there should be no possibility that the employees of a given workplace are also included within the workforce total for some higher level geographical unit. The definition of centroid files requires the specification of several basic parameters relating to the data.

Each centroid has a name (such as "St. George's Hospital") and unique identifier (such as a census OA code or postcode), x and y grid coordinates and a total population count. The total population count is subdivided into a number of sub-groups (e.g. aged 0-3, aged, 4-10, higher education students, etc.) which must sum to the exact total and should have a standard definition within a single data library. The sub-groups are used to differentiate the behaviour of different populations and will be reflected in the available destination data, for example age groups which match divisions in known educational and employment activities. Further parameters are declared for the overall centroid data file as discussed in the following paragraph, but these are default values and may be overridden by exact values provided for any individual centroid.

The name of a timeseries containing profiles of population activity must be declared in a destination centroid file. For example, time profiles of working hours would be associated with workplace centroids. Within the centroid file, subsets of workplaces may be associated with more specific time profiles. A local dispersion function is defined for each centroid, which controls the distance decay algorithm to be used to spread population around the centroid. At present, only the Cressman function used in the original SurfaceBuilder program has been implemented. A local dispersion parameter defines the spatial extent of a centroid – for example a small site like a primary school might be set to 100m, whereas a large site such as a university campus may be set to 1000m. As with the other parameters, the header sets default values which may be overridden if exact information is available in relation to individual centroid records. A wide area dispersion function describes the structure of distance decay in the travel of population to a given destination centroid and is expressed as a series of distance bands with associated percentages of the centroid's population. Additionally, major flows may be explicitly recorded if specific population movements are known to exist between geographical units in the input data, for example a commuting flow from a particular ward to a specific workplace. This structure allows such known population movements to be incorporated into the model before distance decay functions are used to estimate the remaining interactions.

Providing these basic structural requirements are met, the user has considerable flexibility in describing any locations which may at some time contain population. It is advisable to keep all centroids of different types (workplaces, educational establishments, etc.) in separate files to aid data management. Origin and destination centroid files are constructed in comma separated (.csv) format and should conform to the layout described in Annex A for the multi-line header structure which defines various dataset parameters and the field layout of the individual records which follow. In the data block, each centroid is represented by a single line which conforms to the structure defined in the header. For destination centroids, each will need to be associated with a timeseries which indicates the pattern of population occupation of that location relative to its total capacity. More parameters are required for destination than origin centroids. For origin files only, the proportion of each age group which is mobile (i.e. available for spatial reallocation) is defined, allowing "immobile" populations such as prisoners or elderly persons in permanent residential care to be retained at their origin locations.

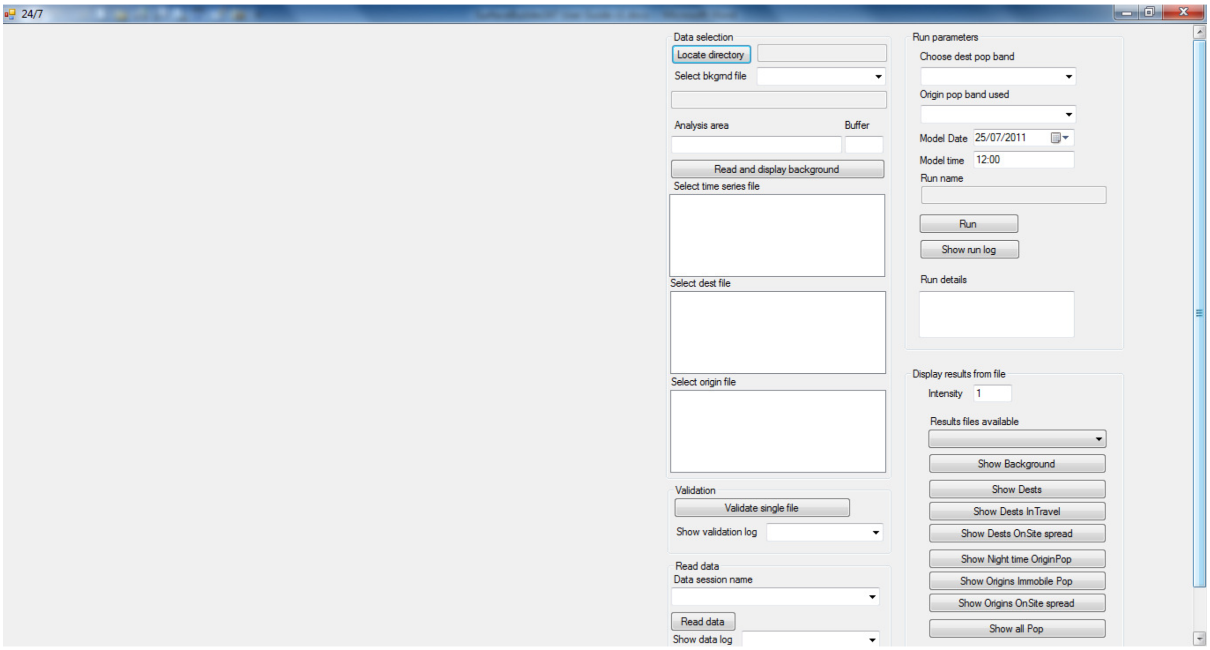
Timeseries is the final essential input data type and takes the form of a time profile library file, in .xls format, which is placed in the timeseries folder with the name timeseries.xls. The timeseries file contains all the time profiles which are identified in the corresponding destination centroid files. The following example shows one simple time profile.

Prim04-11		
InTravel		
	00:00:00	0
	08:30:00	100
	09:00:00	10
	15:00:00	90
	16:00:00	5
	17:00:00	5
	17:30:00	0
OnSite		
	00:00:00	0
	09:00:00	90
	10:00:00	100
	15:00:00	10
	16:00:00	5
	17:00:00	0

A time profile comprises two parts, an InTravel and and an OnSite component, and is always associated with a destination centroid, such that the percentages are expressed relative to the declared population capacity of that destination. The first block represents the times at which given percentages of the destination population are expected to be in the transportation system and the second the times at which they are expected to be at the destination. The example time profile above is called Prim04-11 and describes a day during which 0% of the population in question travel between midnight and 08:30 and 100% travel between 08:30 and 09:00, etc. Any number of time divisions may be used within the day. The second block of time profiles information contains percentages of the capacity population actually present at the destination at different times of day. Further time profiles with variant names may be used to describe the behaviour of the same destination on different days of the week, which might typically be included in different files. Multiple time profiles may be included within a single timeseries file, by adding further pairs of columns. It is not necessary for all the time profiles within a single file to be of the same length.

Running the program

SurfaceBuilder247 should only be run once a complete data library has been assembled as described above. The program can be started by double clicking the STSurfaceBuilderAapp.exe file and will launch an empty modelling workbench, as shown in the figure below.

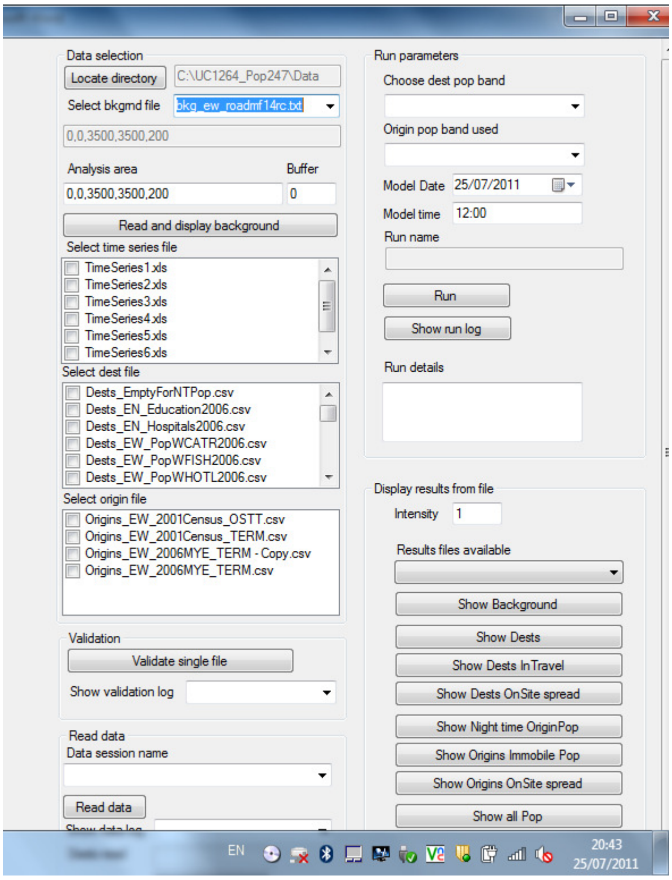


Model Date 25/07/2011
Model time 12:00
Intensity 1

Modelling proceeds by selecting a data library structure, geographical extent and multiple input files within this. The first operation of a program run is to click the “Locate directory” button and navigate to the folder containing the data library. In the standard folder structure described above, this would be C:\UC1264_Pop247 \Data The program will then be able to find the nine sub-folders required for the input and output data files. Once selected, the bckgnd file, time series file list, dest file list and origin file list will all be automatically populated with the names of the relevant files in the selected data library. These files are treated as read-only by the program. In the current version of the software, the next step is to select a background file from the drop-down list. A background file should be chosen which is appropriate to the scenario to be modelled, thus a weighted transportation model suitable for a morning rush hour would look very different to that for a Sunday evening and it is up to the user to select suitable background assumptions. Its geographic extent, dimensions and cell size will then be displayed, reading from the background file header information described above. In the screen shot below a background file bkg_ew_roadmf14rc.txt has been selected, which has xllcorner and yllcorner coordinates of 0, 0. There are 3500 rows and 3500 columns and the cell size is 200m. Note that SurfaceBuilder247 expects all cells to be square.

The analysis area may be adjusted but must not exceed the dimensions of the background file. The analysis area defines the geographical area which will be represented in the results files. A buffer distance is also selected. This is a distance surrounding the analysis area and modelling covers this entire extended area, known as the study area. Input data should cover the entire study area. The study area must be set large enough so that the results in the analysis area are not unduly influenced by population movements from outside the study area, to avoid edge effects. For example, if there is a major settlement 20km beyond the edge of the analysis area which interacts with the analysis area through overlapping travel to work areas, service catchment areas, etc. the buffer should be set greater than 20km. Using UK census data, a buffer of at least 50km is recommended.

At this stage, no other parameters have been chosen and the next stage would be to click the "Read and display background" button. This will produce a map image of the area covered by the background file on the left hand side of the screen. The user now has the option to select multiple time series, dest and orig files by checking the boxes next to their names in the respective lists. It is necessary to choose at least one file from each list.



C:\UC1264_Pop247\Data

Select bkgrnd file: bkg_ew_roadmf14rc.txt

0,0,3500,3500,200

Analysis area: 0,0,3500,3500,200

Buffer: 0

Select time series file

□ TimeSeries1.xls
□ TimeSeries2.xls
□ TimeSeries3.xls
□ TimeSeries4.xls
□ TimeSeries5.xls
□ TimeSeries6.xls

Select dest file

□ Dests_EmptyForNTPop.csv
□ Dests_EN_Education2006.csv
□ Dests_EN_Hospitals2006.csv
□ Dests_EW_PopWCATR2006.csv
□ Dests_EW_PopWFISH2006.csv
□ Dests_EW_PopWHOTL2006.csv

Select origin file

□ Origins_EW_2001Census_OSTT.csv
□ Origins_EW_2001Census_TERM.csv
□ Origins_EW_2006MYE_TERM - Copy.csv
□ Origins_EW_2006MYE_TERM.csv

Data session name

Choose dest pop band

Origin pop band used

Model Date: 25/07/2011

Model time: 12:00

Run name

Intensity: 1

20:43
25/07/2011

Each input file should be validated when first used to ensure that there are no errors in the data structure or integrity. Text format data logs relating to each data file are reported on-screen and also written to the DataLogs folder. As with all the system-generated files, any existing files of the same name will be overwritten. A unique combination of input data files is referred to as a "Data session" and data sessions may be saved and re-read so as to produce precisely the same series of input data in subsequent runs without the need to select all the options manually. The program generates all output file names based on the input data and modelling choices made by the user.

Once an entire, consistent set of input files comprising a background layer, multiple origin and destination centroid files and a timeseries file have been selected, the user is ready to select specific modelling parameters in the right hand column such as the population sub-group to be modelled and the target time and date. The origin and destination population groups available in the dropdown lists are dependent on those defined in the input data files. Once the "Run" button has been clicked, a progress bar tracks the modelling and multiple output files are created from each run, essentially comprising results files, log files and visualization files. Large runs may take several hours and users are strongly advised to test the entire data structure and modelling process using a very small area first: run length is broadly proportional to the number and density of centroids in the study area. The results files, written to the Results subfolder, contain ascii format grid data for further analysis and display files, the latter used only to provide a basic map display within the SurfaceBuilder247 workbench screen. Run logs are written to the RunLog folder and may be pasted into Excel to allow easier analysis.

Summary program operation

During a program run, SurfaceBuilder247 processes all the input centroid data relating to the specified population sub-group within the study area, one centroid at a time, by working through all the specified origin files. Any part of the population sub-group which is not mobile is directly transferred to the same location in the output. The sum of the population counts in these input files defines the population available for modelling. Modelling then proceeds by working through all the destination centroids falling within the study area and considering their time profiles in relation to the target time and date set for the modelling run. Any centroid which has no associated population activity for the sub-group being modelled at the time and date in question is discarded. All remaining centroids therefore have some level of population "demand" at the target time and population counts will be transferred to destinations from origins falling within the wide area dispersion function around each destination. This population is split between those who are OnSite and those InTravel. Thus, the time profile may suggest that a workplace is occupied by 90% of its capacity number of employees of 200 at 09:30. 180 employees from the relevant population sub-group will be transferred away from origin centroids falling within the wide area dispersion function, which is typically derived from census travel to work data or travel survey data. This function will determine what proportions of the 180 employees are obtained from specified distance bands of the destination. If insufficient population supply is available in any band, the search will be widened. Of these 180 employees, the time profile information may suggest that 80% are actually at the place of employment and a further 20% are in the transportation system. The OnSite populations will be assigned to the workplace using the local spread parameter and the 20% in the transportation system will be spread across the background layer using the weights provided by the background file, which will prevent allocation of population into areas such as the sea and will concentrate moving populations onto the transportation network. When all allocations to origins and destinations are complete, the original SurfaceBuilder redistribution algorithm is used to spread these counts across local cells, for example spreading residential populations from a census OA centroid into the surrounding residential areas. The program operates by holding each of these counts in different layers, hence the buttons on the lower right enable the different sub-groups to be displayed on-screen. In order to obtain a model for the entire population, the same modelling scenario must be re-run for each population sub-group and the results summed using suitable GIS software.

Contacts

The project team are continuing to enhance SurfaceBuilder247 and the broader spatio-temporal population modelling project beyond the end of the original ESRC award and are keen to hear from other researchers either intending to use the program independently or interested in contributing to its further development, in which case further details, code and files can be supplied as appropriate. Please register your interest by contacting us:

David Martin D.J.Martin@soton.ac.uk

Samantha Cockings S.Cockings@soton.ac.uk

Geography and Environment, University of Southampton, Southampton, SO17 1BJ, UK

Annex A: SurfaceBuilder247 Origin and destination data file formats

This is a comma separated values file which begins with 20 header records.

Line	Orig	Dest	Keyword	Content
1	Y	Y	Type	Dataset type: “Orig” or “Dest”
2	Y	Y	Title	Dataset title
3	Y	Y	Comment	Comment on data
4	Y	Y	Data block	N1, N2, N3 N1 = row containing data column header N2 = first row of data records N3 = number of data records
5	Y	Y	UniqueID	Data column containing area code (e.g. Postcode, OACode)
6	Y	Y	X	Data column containing X grid coordinate
7	Y	Y	Y	Data column containing Y grid coordinate
8	Y	Y	PopTotal	Data column containing total population for each record
9	Y	Y	PopSubGroups	G1 = Number of population subgroups used
10	Y	Y	PopSubGroupsData	{default values x G1},{columns x G1} {values} = default % values for each of G1 sub-groups (sum = 100) {columns} = data columns containing each population subgroup value
11	N	Y	TimeProfile	Code, N1 Code = Default time profile scheme N1 = Data column containing specific time profiles
12	N	N	LocalDispersionType	Code, N1 Code = Used to define the Local dispersion function – currently not read as Cressman is used by default for all Origins and Dests (with spatial extent as defined in row 12 (LocalDispersion) N1 = column defining centroid-specific LocalDispersion Type, if any
13	N	Y	LocalDispersionParameter	N1, N2 N1 = Default spatial extent of locations N2 = Data column containing specific spatial extents
14	N	Y	WideAreaDispersion	Spatial range distribution, N1 e.g. 100>5|2000>80|5000>5|10000>3|20000>2|5 where: 5% pop work from home (narrow radius of 100 specified), 80% pop >100 to 2000m, 5% >2000-5000m, 3% >5000 to 10000m and 2% >10000 to 20000 and finally 5% > 20000 – the radius of the final band is based on a uniform distribution density of all populations (in above example, 95% is within 20000, so radius of final band is 20000/sqrt(.95) N1 = Data column containing spatial range distribution
15	N	Y	MajorFlows	{Columns x G} – only for Dests e.g. 00MSNE>5|00MSMY>5|00MSNA>2 – this represents 3 major flows for the dest – 5% of its population comes from origins with IdentifyingCode begining with the characters ‘00MSNE’, 5% from origins with IdentifyingCode begining with the characters ‘00MSMY’ and 2% from origins with IdentifyingCode begining with the characters ‘00MSNA’. Notes that only origins within the study area are included.
16	Y	Y	DOA	Reference date, N1 Reference date = textual description of reference data

				N1 = Data column containing specific reference dates
17	Y	N	Mobility	{default values x G1},{columns x G1} Only for Origins Mobility is a percentage and must be between 0 and 100 {default values} = Mobility for each PopSubGroup {columns} = data columns containing each population sub-group value
18	Y	N	RegionalIdentifier	Default region code (e.g. UK), N1 Only for Origins N1 = Data column containing specific RegionalIdentifiers (e.g. SW, SE)
19	Y	N	RegionalAdj	{values x G} Only for Origins A vector of RegionalIdentifiers and percentages (e.g. SW>109|SE>96|UK>100 = original value*109% in Southwest; original value*96% in Southeast; original value*100% in UK)
20	Blank line			
21	Blank line			

Line 22 (defined as N1 in data block field above) contains column numbers for ease of reference

Line 23 (defined as N2 in data block field above) is the first data record

Data block continues for N3 lines (as defined in data block field above)