Spatial Allocator Surrogate Tool: User's Guide

Limei Ran
Carolina Environmental Program
The University of North Carolina at Chapel Hill
137 E. Franklin St., CB 6116
Chapel Hill, NC 27599-6116

Updated: May 30, 2014

Contents

1. Introduction

This document is a User's Guide for the Surrogate Tool, which is a stand-alone Java tool for generating spatial surrogates that are inputs to emission models such as the Sparse Matrix Operator Kernel Emissions (SMOKE) modeling system in support of Eulerian grid models. The Surrogate Tool, which is implemented in EPA's Emissions Modeling Framework (EMF), is a component of the Spatial Allocator (SA) system. One goal of the EMF is to make it easier to produce, maintain, and track SMOKE ancillary files, and the Surrogate Tool with SA helps to accomplish this goal. The Surrogate Tool uses user-defined text inputs to control which surrogates are generated, and the format of these files allows them to be easily edited and maintained in a spreadsheet program like Microsoft Excel. When the Surrogate Tool is used from the EMF, the EMF Data Management system provides a graphical user interface (GUI) for users to store, edit, and manage their Surrogate Tool input and output files.

The Surrogate Tool is built upon the Spatial Allocator Vector Tools (http://www.cmascenter.org/sa-tools/documentation/4.0/html/vector_tools.html), which has the features needed to produce spatial surrogates, but it can be complex to use. The Spatial Allocator tools were developed over several years and are run using C-shell scripts on UNIX . The tools are C programs that must be compiled on each operating system. Currently, they are released for Linux. They have also been used on other UNIX platforms, such as Solaris, IRIX, and AIX. Additional information is available in the Spatial Allocator documentation, available on the website listed above.

The Surrogate Tools are now packaged as part of the Spatial Allocator Release, installation instructions are provided below.

The Surrogate Tool was developed to provide a more user-friendly way to use the Spatial Allocator. Because it has been developed using Java, the Surrogate Tool can generate surrogates regardless of the operating system used, as long as the Spatial Allocator programs can be compiled on that operating system. Unlike the Spatial Allocator alone, users do not need to define environment variables, create intermediate text files, or use scripts. Users define all information needed to generate surrogates using ASCII files. In addition, the Spatial Allocator can now generate surrogates for polygons (e.g., census tracts to be used by the ASPEN dispersion model) and E-Grids used in the new WRF/NMM-CMAQ system. The Surrogate Tool interfaces with the Spatial Allocator to support the generation of grid, E-Grid, and polygon based spatial surrogates. The Surrogate Tool uses Comma Separated Value (CSV) files that define how the surrogates should be computed, what data should be used, where the results should be stored, which surrogates should be computed during the run, and other needed information. The Surrogate Tool also reads in a list of ESRI Geographic Information System (GIS) shapefile names and their map projection information for the Spatial Allocator from a file called a "shapefile catalog". The Surrogate Tool takes a single command line argument and can be run using interactive or batch mode. After it is run, users can check a log file to make sure that no errors have been reported by the Tool.

The output from the Surrogate Tool is a set of SMOKE spatial surrogate ratio files that can be input to SMOKE for grid-based (either regular or E-Grid), or polygon-based modeling. Although SMOKE version 2.3 does not yet support the use of polygon-based surrogates. Each

surrogate created by the tool is placed into an individual file, and a concatenated file of all surrogates can also be created if requested. Note that SMOKE versions 2.3 and higher support using spatial surrogates from separate files (one file per surrogate); but SMOKE versions 2.2 and lower expect all surrogates to be in a single file. Under the newer approach, the multiple surrogate files are listed in the Surrogate Description (SRGDESC) file so that SMOKE can find all of the available surrogates.

2. Installing the Spatial Allocator (contains the Surrogate Tool)

To install the Spatial Allocator, follow the instruction in the Spatial Allocator documentation.

3. Using the Surrogate Tool

3.1 Input files

The Surrogate Tool makes generating spatial surrogates easier than using the Spatial Allocator software alone. To create the desired surrogates, the Surrogate Tool runs the Spatial Allocator program "srgcreate" to generate surrogates directly from shapefiles, and also performs surrogate merging and gapfilling. The Surrogate Tool's input files are five comma-separated-value (.CSV) files and a grid description file. Each CSV file is a tabular file that requires a specific set of columns. The title of each column describes the meaning of the data in the column and also notifies the Tool of the contents of the column. These CSV files can easily be viewed and edited by any spreadsheet software. Examples of each of the input files are provided with the Surrogate Tool installation in the $SA_HOME/srgtools directory and can be customized to meet your needs. Detailed descriptions of each of the input files are in the subsections of this section. The high-level descriptions of the input files are as follows:

1. The global "control variables" file is a CSV file that specifies information common to the generation of all surrogates (e.g., output directory, output grid or polygons, and names of the other input files). The sample file names are control_variables_grid.csv for a regular grid example, control_variables_egrid.csv for an egrid example, and control_variables_poly.csv for a polygon-based example. The variables that should be set in the global control variables file are described in detail in the section 3.2.

2. The shapefile catalog file is a CSV file that provides location, map projection and source information about the shapefiles to be used during surrogate generation. The sample file name is shapefile_catalog.csv. The content of this file is described in more detail in section 3.3.

3. The surrogate specification file (SSF) is a CSV file that provides information needed to generate each surrogate, including the input shapefiles or previously computed surrogates, weight attributes or merge functions to use, shapefile filter selections to apply, and how the surrogates should be gap-filled. The sample file name is

surrogate_specification_2002.csv. The content of this file is described in more detail in section 3.4.

4. The surrogate code file is a CSV file that provides surrogate names and codes that are used to map surrogate names to surrogate codes, which is needed during surrogate merging and gapfilling. The sample file name is surrogate_codes.csv. The content of this file is described in more detail in section 3.5.

5. The generation control file is a CSV file that specifies the surrogates to create for a specific run of the Surrogate Tool and whether to output quality assurance data for those surrogates (i.e., numerators, denominators, and sums of fractions for the county). The sample file name is surrogate_generation_grid.csv. The content of this file is described in more detail in section 3.6.

6. The grid description file is a text file that provides grid description for a grid name. The sample file included is GRIDDESC.txt. The grid used in the sample is named "US36KM_148X112". Users can add new grid name and grid description to this file for their own computation. For more information on the format of the GRIDDESC file, see http://www.baronams.com/products/ioapi/GRIDDESC.html.

3.2 Global Control Variables File

The global control variables file is a CSV file that specifies information that is common to the generation of all surrogates (e.g., output directory, output grid or polygons, and names of the other input files). A single global control variables file is used for each run of the Surrogate Tool. The columns VARIABLE and VALUE are required. Any additional columns are optional and are ignored by the Surrogate Tool. There are a number of variables that should be set in the global control variables file. The contents of the global control variables file for the RegularGrid output type are shown in Tables 1 and 2. (Tables 1 through 5 are shown together following this discussion). Table 1 shows the file as it would appear loaded into a spreadsheet. Table 2 shows the file as it would appear loaded into a standard text editor. The following variables (listed in capital letters below) are recognized by the Surrogate Tool in the global control variables file:

- GENERATION CONTROL FILE gives the directory and name of the generation control CSV file to use for the run.
- SURROGATE SPECIFICATION FILE gives the directory and name of the surrogate specification CSV file to use for the run.
- SHAPEFILE CATALOG gives the directory and name of the shapefile catalog CSV file to use for the run.
- SHAPEFILE DIRECTORY gives the name of the directory in which to look for the shapefiles in the shapefile catalog.
- SURROGATE CODE FILE gives the directory and name of the surrogate code CSV file to use for the run.
- SRGCREATE EXECUTABLE gives the directory and name of the srgcreate executable to use for the run.

- DEBUG_OUTPUT specifies whether srgcreate will output debugging information as it runs (specify Y for yes and N for no).
- OUTPUT_FORMAT specifies the format for the output files (currently SMOKE is the only allowable value).
- OUTPUT_FILE_TYPE specifies the type of output file to create (currently RegularGrid, EGrid, and Polygon are the allowable values). The RegularGrid option should be used to generate surrogates for Eulerian grid-based models such as CMAQ. EGrid should be used only for the WRF/NMM-CMAQ system, whereas the Polygon option is used for non-grid-based models such as ASPEN.
- OUTPUT_GRID_NAME specifies the name of the output grid (valid only when OUTPUT_FILE_TYPE is RegularGrid or EGrid).
- GRIDDESC specifies the directory and name of the grid description file (valid only when OUTPUT_FILE_TYPE is RegularGrid or EGrid).
- OUTPUT_FILE_ELLIPSOID specifies the ellipsoid of the output grid (valid only when OUTPUT_FILE_TYPE is RegularGrid or EGrid).
- OUTPUT_POLY_FILE specifies the name of an ArcGIS polygon text file or the name of a shapefile containing the polygon shapes to use (valid only when OUTPUT_FILE_TYPE is EGrid or Polygon).
- OUTPUT_POLY_ATTR specifies the name of the attribute in the OUTPUT_POLY_FILE that is a unique ID for each shape (valid only when OUTPUT_FILE_TYPE is Polygon).
- OUTPUT DIRECTORY specifies the name of the directory into which the output surrogate files will be placed.
- OUTPUT SURROGATE FILE specifies the name of the optional file that combines all of the surrogates created during the run into a single file (this is not needed with version 2.3 and higher of SMOKE, but is used to support earlier versions). If this variable is defined, the combined single file will be created; otherwise, it will not be created. This file is placed in the same directory as the individual surrogate files.
- OUTPUT SRGDESC FILE specifies the directory and name of the output surrogate description file (SRGDESC file) that is used as an input to SMOKE.
- OVERWRITE OUTPUT FILES specifies whether to overwrite output files if they exist (YES or NO are the allowable values). If this is set to NO and the output files already exist, the Surrogate Tool will end with an error. If this is set to YES and the output files already exist, the output files will be overwritten.
- LOG FILE NAME specifies the directory and name (full path) of the Surrogate Tool log file.
- DENOMINATOR_THRESHOLD specifies the value of a threshold under which the surrogate values will not be used (but may be replaced with a gap-filled value, if gap filling is used). The default value is 0.00001. Denominators of this size occur when the intersected county and weight polygons are tiny (e.g., they are both for county data and the lines do not exactly line up). This is explained in more detail below. If users do not wish to use the

denominator threshold feature when writing the surrogates, the value of this variable should be set to 0.0

- COMPUTE SURROGATES FROM SHAPEFILES specifies whether or not this run of the Surrogate Tool will compute surrogates from shapefiles. If it is set toYES, the Surrogate Tool will compute surrogates from surrogate shapefiles by calling srgcreate.exe of the Spatial Allocator based on the contents of the surrogate specification file.
- MERGE SURROGATES specifies whether or not this run of the Surrogate Tool will compute surrogates by merging existing surrogates using the merging tool. If it is set to YES, the run will compute surrogates from the merging tool as specified in the surrogate specification file.
- GAPFILL SURROGATES specifies whether or not this run of the Surrogate Tool will gapfill existing surrogates using the gapfilling tool. If it is set to YES, the run will gapfill surrogates as specified in the surrogate specification file.

These variables can be specified in any order, one per line. The Tool writes a warning to the log file if there are unrecognized variable names. Users can customize the sample control CSV files that are provided with the Surrogate Tool for their application: the sample file "control_variables_grid.csv" is for regular-grid-based surrogates, "control_variables_egrid.csv" is for egrid-based surrogates, and "control_variables_poly.csv" is for polygon-based surrogates.

The variable DENOMINATOR_THRESHOLD is used to prevent surrogates from being output for tiny areas that result from offsets of the same boundaries that appear in both data and weight shapefiles due to different data sources and processes (e.g. county data versus population data comprised of census tracts). The numerator for a surrogate ratio is equal to the surrogate weight value in the intersected region of a modeling polygon (i.e., a grid cell in a RegularGrid or EGrid, or a polygon) and a base data polygon (e.g., county) and the denominator is the total surrogate weight value for the entire base data polygon. The Surrogate Tool runs the srgcreate program which overlays the modeling polygons (such as grids) with the base data polygons (such as county polygons) and weight shapes (such a census tracts for population, roads, or airports) to perform the surrogate ratio computation. If the denominator is smaller than the specified DENOMINATOR_THRESHOLD, the surrogate ratio is output as a comment line starting with # (which causes the line to be ignored by SMOKE and the surrogate merging and gapfilling tools).

Note that the denominator threshold is not a ratio, it is an absolute value. Thus the size of the attributes being allocated needs to be considered when setting this value. Typically population and most attributes used for current modeling have values substantially more than 1, so setting the threshold on the order of 1E-5 is reasonable. In our runs, we found it was useful to set this value to 0.0005 to prevent weight data from showing up in a county which does not really have any surrogate weight value. Since the denominator is the total surrogate value for entire data polygon such as a county, all grids intersecting the data polygon will then appear as comment lines. For quality assurance purposes, a comment line is added to the output surrogate file which specifies the surrogate code, county ID, 0 for the row and column and the residual ratio for the county if this the surrogate ratios do not sum to 1 for a county due. This may occur if the data polygon extends beyond the grid domain, due to numerical rounding in calculation or for any other reason. **Note that all comment lines in the original surrogate data are lost in merged**

or gapfilled surrogate output files because the merging and gapfilling tools remove any lines starting with a '#'.

An example of the problem that the DENOMINATOR_THRESHOLD prevents is as follows. During development of the Surrogate Tool (prior to implementing this feature), we observed that surrogate ratios were computed for counties that had very small values for the denominators. This occurred in particular for surrogates computed from weight shapefiles containing county or census tract boundaries such as EPA's Urban Population (surrogate code 120). For example, the urban population surrogate might have a denominator for a particular county of 3.5×10^{-5}. If the county boundaries in the weight shapefile did not exactly match those in the data shapefile, then some spatial artifacts were occurring. These artifacts caused some counties with zero values for the urban population attribute in the census population shapefile to appear with nonzero values in the surrogate output file as a result of a tiny contribution (spatial artifact) from an adjacent county. This was the result of the county boundaries not matching exactly in the two files so that small portions of counties with non-zero urban population would be allocated to adjacent counties with zero urban populations. As a result, surrogate values would be output for those counties. This type of problem can now be eliminated using the DENOMINATOR_THRESHOLD variable.

3.3 Shapefile Catalog

The shapefile catalog file is a CSV file that provides location, projection and source information about the shapefiles to be used during surrogate generation. You can specify the path for all shapefiles in the control variable CSV file using SHAPEFILE DIRECTORY variable. If the variable is not defined in the control file, the path specified in the shapefile catalog will be used. The sample shapefile catalog provided with the Surrogate Tool is shapefile_catalog.csv. The shapefiles used by this catalog are available as a gzipped tar file (emiss_shp2003.tar.gz), that will be installed under the $SA_HOME/data/emiss_shp2003 directory. This data includes US shapefiles. You can download additional shapefiles for Canada from the EPA Emission Inventory web site (ftp://ftp.epa.gov/EmisInventory/emiss_shp2003/). Check the SHAPEFILE DIRECTORY value in the control variable CSV file and any relative paths in the shapefile catalog file to verify that they are consistent with the locations on your computer.

An example of a shapefile catalog file as it would look loaded into a spreadsheet is shown in Table 3. Note that this file could also be edited using a standard text editor, but that view of it is not shown here. There are four columns that must be specified for each line of the shapefile catalog: SHAPEFILE NAME, DIRECTORY, ELLIPSOID, and MAP PROJECTION. Any subsequent columns are optional and are ignored by the Surrogate Tool. Note that the entries in the DIRECTORY column can be used to specify relative paths beneath your main SHAPEFILE_DIRECTORY, which is specified in the control variable file. Some recommended additional columns for metadata purposes are SHAPE TYPE (point, line, or polygon), DESCRIPTION, DATA SOURCE (the source of the shapefile), RESOLUTION (the level of detail of the shapefile), DATE OF DATA (the date to which the data applies), and RETRIEVAL DATE (the date the file was obtained).

The ELLIPSOID and MAP PROJECTION columns should follow the syntax specifications of the Spatial Allocator, which passes this information on to the PROJ.4 library (see

http://www.ie.unc.edu/cempd/projects/mims/spatial/ for more information). It is important to note that SHAPEFILE NAME must be unique in the catalog. The names are critical because they are how the Surrogate Tool obtains the location and map projection information of the shapefiles used to create the surrogates. It is essential that these shapefile names be consistent with those used in the surrogate specification file.

3.4 Surrogate Specification File

The surrogate specification file (SSF) is a CSV file that provides information needed to generate each surrogate. This includes the input shapefiles or previously computed surrogates, weight attributes or merge functions to use, shape filters to apply, and how the surrogates should be gap-filled. The value of the SURROGATE SPECIFICATION FILE variable in the global control variables file sets the file location and name of the SSF that the Surrogate Tool will use during a given run. An example surrogate specification file as it would look loaded into a spreadsheet is shown in Tables 4a and 4b (note that the format of each row is split into parts (a) and (b) so that the information fits on the pages of this document). The sample specification file provided with the tool – is named "surrogate_specification_2002.csv". It can be modified by adding new surrogates to the table or by changing the attributes used for surrogate computation. This file can be edited using a standard text editor; however, due to the large number of columns, we believe that a spreadsheet program is a better choice for editing the file.

The SSF contains 13 columns that are recognized by the Surrogate Tool. Any additional columns are optional and are ignored. The recognized columns are:

1. REGION: the name of the region for the surrogate (e.g., USA, Canada).

2. SURROGATE: the name of the surrogate to create (e.g., Population, Water).

3. SURROGATE CODE: the code number used for the surrogate. Note that the combination of REGION and SURROGATE CODE must be unique in the SSF.

4. DATA SHAPEFILE: the name shapefile to use for the base [data] polygons (e.g., counties, provinces). The name of this shapefile must appear in the SHAPEFILE NAME column of the shapefile catalog.

5. DATA ATTRIBUTE: the attribute to use to create the surrogate from a shapefile. This is not used if this surrogate is being created by merging existing surrogates. **Note: surrogates generated for any data polygons that do not have a value for the data attribute will be written as comments to the output files by the srgcreate program. They will not be preserved by the surrogate merging and gapfilling tools because all intermediate comment lines will be removed during merging or gapfilling.**

6. WEIGHT SHAPEFILE: the name of the shapefile used for the weight shapes (e.g., census tracts, railroad lines, port points). This is not used if this surrogate is being created by merging existing surrogates. The name of this shapefile must appear in the SHAPEFILE NAME column of the shapefile catalog.

7. WEIGHT ATTRIBUTE: the name of the attribute to use for computing the weights of the surrogate (e.g. POP2000, BERTHS). Specify NONE to use the area for polygons, length

for lines, or point counts for points. This is not used if this surrogate is being created by merging existing surrogates, or if a WEIGHT FUNCTION is being used.

8. WEIGHT FUNCTION: a function to use for computing the weight of the surrogate. The attributes used in the function should exist in the weight shapefile. The weight function can be any arithmetic equation containing the operators +, -, *, /, (,), numeric constants, and names of attributes that exist in the weight shapefile. Exponential notation and power functions are not currently supported, nor are unary negative numbers used as constants (e.g., X1 + -5 should be X1 - 5). Examples of acceptable weight functions are: WEIGHT_FUNCTION=(IND1+IND2+IND3+IND4+IND5) or WEIGHT_FUNCTION=0.75*urban+0.25*rural (see http://www.ie.unc.edu/cempd/projects/mims/spatial/weight_func.html for more information).

9. FILTER FUNCTION: specifies "filter" or selection criteria for shapes to include or not include in the surrogate computation (e.g., ROAD_TYPE!=2) excludes all shapes for which ROAD_TYPE does not equal 2, and GRID_CODE=61,81,82 includes all shapes for which the GRID_CODE is 61, 81, or 82). Multiple filters can be specified if they are separated by semicolons (e.g., LENGTH=100-200;NAME=C*). This function is not used if this surrogate is being created by merging existing surrogates **(see http://www.ie.unc.edu/cempd/projects/mims/spatial/filters.html for more information about the filtering syntax).**

10. MERGE FUNCTION: specifies a function to use when creating a surrogate by merging or concatenating existing surrogates. Referenced surrogates can be in the SSF or external (e.g., 0.5*../data/surrogate_file|Forest+0.5*Rural Land, Population[US];Population[Canada],Population[Mexico]) where the '|' character separates the name of the file containing the surrogate(s) from the name of the surrogate itself, and the string within the brackets corresponds to a region name. A further description of the syntax is given below.

11. SECONDARY SURROGATE: the name of a surrogate to use as a secondary surrogate to gapfill the values of the primary surrogate. Referenced surrogates can be in the SSF or external (e.g., Population, ../data/surrogate_file|surrogate_name).

12. TERTIARY SURROGATE: the name of a surrogate to use as a tertiary surrogate to gapfill the values of the primary surrogate. Referenced surrogates can be in the SSF or external (e.g., Population[Mexico], ../data/surrogate_file|surrogate_name).

13. QUARTERNARY SURROGATE: the name of a surrogate to use as a quarternary surrogate to gapfill the values of the primary surrogate. Referenced surrogates can be in the SSF or external (e.g., Population[Mexico], ../data/surrogate_file|surrogate_name).

Recall that the combination of REGION and SURROGATE CODE values must be unique in the SSF. For example, you may wish to generate population surrogates with the same surrogate code 100 for the regions USA, Canada, or Mexico. To do this, you can specify one row for each region, but on each row use the surrogate code 100. The SMOKE version 2.3 and higher

supports reading all surrogates with the same code from the SRGDESC file, so the surrogate files for each region do not need to be concatenated.

For surrogates generated directly from shapefiles, the DATA SHAPEFILE column specifies the name of the base polygons for emission sources, such as county, census tract, or other polygons. The DATA ATTRIBUTE column specifies the name of the attribute to uniquely identify the base polygons (e.g., county FIPS code). The WEIGHT SHAPEFILE column specifies the name of the weight (surrogate) shapefile for surrogate ratio computation, such as population, road, or land use shapefiles. The WEIGHT ATTRIBUTE column specifies the name of the attribute to use for the surrogate computation (e.g., population). When the WEIGHT ATTRIBUTE is specified as NONE, the value input as the weight for a shape is its area for polygon weight shapefile, length for line weight shapefile, or point count for a point weight shapefile.

If a function of multiple attributes is to be used for the weight, this is specified in the WEIGHT_FUNCTION column (e.g., COM1+COM2+COM3). In cases where not all shapes from the shapefile are to be used to generate the surrogate, a FILTER FUNCTION is specified (e.g., ROAD_TYPE=1,2,3 to use only shapes with road types of 1, 2, or 3; or ROAD_TYPE != 1 with road type not equal to 1). Multiple filters can be specified if they are separated by semicolons (e.g., LENGTH=100-200;NAME=C*).

Gap filling will be performed if surrogates are given in the SECONDARY SURROGATE, TERTIARY SURROGATE, or QUARTERNARY SURROGATE columns. Gap filling is used when a surrogate does not have values for a base data polygon in the modeling domain. A county will not have any surrogate ratios when the value of the weight attributes for the county are zero, there are no weight shapes that intersect the county, or the total weight surrogate of this county (denominator in surrogate ratio computation) is less than DENOMINATOR_THRESHOLD). Gapfilling ensures that every county with emission inventory data has the surrogate ratios to distribute the emission data. For example, the inventory could have railroad emissions in a county, even if the weight shapefile used to create a railroad surrogate did not have data in that county for any railroads. In this case, the roads surrogate could be used as the secondary surrogate.

If the surrogate to be computed is a function of other surrogates, a MERGE FUNCTION should be specified (e.g., 0.75*Roadway Miles + 0.25*Population). Careful consideration needs to be given regarding how to gapfill surrogates that use a merge function. This is because when merging, the srgmerge program does not output values for any counties that do not have values for all surrogates that are referenced in the merge function. To extend the 0.75 Total Roadway Miles plus 0.25 Population surrogate example, if Total Roadway Miles were missing for a particular county, srgmerge cannot know that the solution is to use 1 * Population. You can account for this by gapfilling your merged surrogate with the input surrogates in the order that you prefer (e.g., you might gapfill the 0.75 Total Roadway Miles plus 0.25 Population surrogate with Total Roadway Miles and then Population).

Surrogates can be concatenated into a single output file by writing a MERGE FUNCTION that has the individual surrogates separated with semicolons. If the region for the source surrogates is different from the region of the output surrogate, the syntax: *surrogate*[*region*] is used. Note that the headers for the concatenated surrogates will appear at the top of each surrogate. An

example of concatenation is to merge population surrogates from North America. To do this, one would use the following syntax in the MERGE FUNCTION column:

Population[USA];Population[MEXICO];Population[CANADA].

External surrogates can be specified as input for merging or gap filling using the following syntax: *file name* | *surrogate name*. If the merging and gapfilling tools are updated to accept codes in addition to names in its input file, the syntax*: file name* | *surrogate code* will also be supported. Until that time, the surrogate names should be specified in the surrogate code CSV file using the syntax:

#SRGDESC=*surrogate code, surrogate name*

For example, you might have the following records in your surrogate file:

#SRGDESC=100,Population
#SRGDESC=110,Housing
#SRGDESC=120,Half population half housing

3.5 Surrogate Code File

A surrogate code file is a CSV file used by surrogate merging and gapfilling tools that specifies the mapping of surrogates names to surrogate codes. This is required because merging and gapfilling use the names of surrogates in their text input files. The syntax of this file is just a collection of #SRGDESC lines, as shown at the end of the preceding section. The Surrogate Tool will find surrogate codes from this CSV file using the surrogate names. The sample surrogate code CSV file named "surrogate_codes.csv" is included with the Surrogate Tool and contains surrogate names and codes from 100 to 940 used by the US EPA for the USA and Canada regions. When external surrogates are used in merge and gapfill functions, users need to add external surrogate entries to the CSV file.

3.6 Generation Control File

The generation control file is a CSV file that specifies the surrogates to create for a specific run of the Surrogate Tool. Users can modify the sample generation control file provided with the Surrogate Tool, named "surrogate_generation_grid.csv", for their computation (see Table 5). The columns REGION, SURROGATE, SURROGATE CODE, GENERATE, and QUALITY ASSURANCE are required to be included in the file. If the value in the GENERATE column is YES, the surrogate will be generated. If the value in the QUALITY ASSURANCE column is YES, surrogate ratios will be output with the numerator, denominator, and quality assurance sum for each surrogate fraction. The quality assurance sum is a running total of the sum of the surrogate fractions for a particular base data polygon (e.g., county). Rows must exist in the surrogate specification file with the same values for the REGION, SURROGATE, and SURROGATE CODE columns. To ensure consistency, you may wish to copy these columns directly from the surrogate specification file and paste them into this file to create it.

Table 1. Example of a Global Control Variables File for RegularGrid Loaded into a Spreadsheet (control_variables_grid.csv)

VARIABLE	VALUE	DESCRIPTION
GENERATION CONTROL FILE	./surrogate_generation_grid.csv	File containing surrogates for computation
SURROGATE SPECIFICATION FILE	./surrogate_specification_2002.csv	File containing settings for generating surrogates
SHAPEFILE CATALOG	./shapefile_catalog.csv	Shapefile names and map projection information
SHAPEFILE DIRECTORY	../data/emiss_shp2003/us	Directory containing all shapefiles needed
SURROGATE CODE FILE	./surrogate_codes.csv	List of surrogate codes and names
SRGCREATE EXECUTABLE	../bin/srgcreate.exe	Location of srgcreate executable
SRGMERGE EXECUTABLE	Java	(use Java merge and gapfill)
DEBUG_OUTPUT	Y	Output debug control
OUTPUT_FORMAT	SMOKE	Output files used for SMOKE
OUTPUT_FILE_TYPE	RegularGrid	Type of output shapes being generated - RegularGrid or Polygon
OUTPUT_GRID_NAME	M08_NASH	This is a grid name for regular grid output area
GRIDDESC	./GRIDDESC.txt	It is the file containing the list of available of grids (needed only for SMOKE surrogates)
OUTPUT_FILE_ELLIPSOID	"+a=6370000.0,+b=6370000.0"	Output grid projection ellipsoid
OUTPUT DIRECTORY	../output/M08_NASH	Directory for individual surrogate files
OUTPUT SURROGATE FILE	../output/M08_NASH/srg_total.txt	Name and path for the final merged surrogate file output from srgtool
OUTPUT SRGDESC FILE	../output/M08_NASH/SRGDESC.txt	File with surrogate codes and description
OVERWRITE OUTPUT FILES	YES	Users can choose YES to overwrite the individual and total output surrogate ratio files
LOG FILE NAME	srg_grid.log	Log file to store all information from running the program
DENOMINATOR_THRESHOLD	0.0005	Surrogate ratio is output as comment line with # sign if denominator of surrogate ratio computation is less than the threshold (default=1E-5)
COMPUTE SURROGATES FROM SHAPEFILES	YES	If set to YES, srgcreate is called to compute surrogates
MERGE SURROGATES	YES	If set to YES the surrogates will be merged
GAPFILL SURROGATES	YES	If set toYES, the surrogates will be gapfilled

Table 2. The Global Control Variables File for RegularGrid as A CSV File (control_variables_grid.csv)

```
VARIABLE,VALUE,DESCRIPTION
GENERATION CONTROL FILE,./surrogate_generation_grid.csv,File containing surrogates for computation
SURROGATE SPECIFICATION FILE,./surrogate_specification_2002.csv,File containing settings for generating surrogates
SHAPEFILE CATALOG,./shapefile_catalog.csv,Shapefile names and map projection information
SHAPEFILE DIRECTORY,../data /emiss_shp2003/us,Directory containing all shapefiles needed
SURROGATE CODE FILE,./surrogate_IDs.csv,List of surrogate codes and names
SRGCREATE EXECUTABLE,../bin/srgcreate.exe,Location of srgcreate executable
SRGMERGE EXECUTABLE,Java, set to Java to use Java gapfilling and merging
DEBUG_OUTPUT,Y,Output debug control
OUTPUT_FORMAT,SMOKE,output files used for SMOKE
OUTPUT_FILE_TYPE,RegularGrid,Type of output shapes being generated - RegularGrid or Polygon
OUTPUT_GRID_NAME,M08_NASH,"This is a grid name for output area."
GRIDDESC,./GRIDDESC.txt,"It is the file containing the list of available of grids (needed only for SMOKE surrogates)."
OUTPUT_FILE_ELLIPSOID, "+a=6370000.0,+b=6370000.0","Output grid projection ellipsoid for the grid."
OUTPUT DIRECTORY,../output/M08_NASH,Directory for individual surrogate files
OUTPUT SURROGATE FILE,../output/M08_NASH/srg_total.txt,name and path for the final merged surrogate file output from
srgtool
OUTPUT SRGDESC FILE,../output/M08_NASH/SRGDESC.txt,file with surrogate codes and description
OVERWRITE OUTPUT FILES,YES,Users can choose YES to overwrite the individual and total output surrogate ratio files
LOG FILE NAME,srg_grid.log,log file to store all information from running the program
DENOMINATOR_THRESHOLD,0.0005,Surrogate ratio is output as comment line with # sign if denominator of surrogate ratio
computation is less than the threshold
COMPUTE SURROGATES FROM SHAPEFILES,YES,"If set to YES, srgcreate is called to compute surrogates."
MERGE SURROGATES,YES," If set to YES, the surrogates will be merged."
GAPFILL SURROGATES,YES," If set to YES, gapfilling will be performed."
```

Table 3. Example of a Shapefile Catalog Loaded into a Spreadsheet (shapefile_catalog.csv).

SHAPEFILE NAME	DIRECTORY	ELLIPSOID	PROJECTION	SHAPE TYPE	DESCRIPTION	DATA SOURCE
county_pophu02	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	US county polygon data from shapefile pophu2k	Extracted and edited from pophu2k
county_pophu02water	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	US county polygon data from shapefile pophu2k	Extracted and processed from pophu2k
pophu2k	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	Population and housing units from Census 2000	US Census Bureau
vi_pophu2k	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	Population and housing units from Census 2000 for Virginia Islands	
us_ph	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	The change in housing between 1990 and 2000	Computed
us_heat	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	Number of housing units in primary heating categories for each census block	US Census Bureau
usrds_2000	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	Proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Line	primary and secondary roads for urban and rural areas	US Census Bureau – TIGER
us_rail2k	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Line	Class 1-3 and unknown	Transportation Atlas Data &

					classified railroads	Census 2000 TIGER data
us_lowres	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	Area of NLCD Low Intensity Residential Land	NLCD
us_ag2k	../data/emiss_shp2003/us	+a=6370997.0,+b=6370997.0	proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Polygon	Agricultural lands—areas of Pasture/Hay, Grains, Row Crops, Fallow Land and Orchards/Vineyards	NLCD

Table 4a. Example of the Left Columns of the Surrogate Specification File Loaded into a Spreadsheet (surrogate_specification_2002.csv)

REGION	SURROGATE	SURROGATE CODE	DATA SHAPEFILE	DATA ATTRIBUTE	WEIGHT SHAPEFILE	WEIGHT ATTRIBUTE	WEIGHT FUNCTION	FILTER FUNCTION
USA	Population	100	county_pophu02	FIPSSTCO	pophu2k	POP2000		
USA	Urban Population	120	county_pophu02	FIPSSTCO	pophu2k	URBAN		
USA	Residential Heating - Natural Gas	150	county_pophu02	FIPSSTCO	us_heat	UTIL_GAS		
USA	Total Road Miles	240	county_pophu02	FIPSSTCO	usrds_2000	NONE	_	
USA	Urban Primary Road Miles	200	county_pophu02	FIPSSTCO	usrds_2000	NONE		NEWRD_CLAS = 1
USA	0.75 Total Roadway Miles plus 0.25 Population	255						
USA	Land	340	county_pophu02	FIPSSTCO	us_lw2k	NONE		H20_CODE=2
USA	Water	350	county_pophu02water	FIPSSTCO	us_lw2k	NONE		H20_CODE!=2

USA	Rural Land Area	400	county_pophu02	FIPSSTCO	rural_land	NONE		RL_FLAG=Rural Land
USA	Total Agriculture	310	county_pophu02	FIPSSTCO	us_ag2k	NONE		GRID_CODE=61,81, 82,83,84
USA	Industrial Land	505	county_pophu02	FIPSSTCO	us_lu2k		IND1+IND2+ IND3+IND4+ IND5+IND6	
USA	Heavy and High Tech Industrial (IND1 + IND5)	570	county_pophu02	FIPSSTCO	us_lu2k		IND1+IND5	
USA	Forest External	328						
NA	Population	100						

Table 4b. Example of the Right Columns of the Surrogate Specification File Loaded into a Spreadsheet (surrogate_specification_2002.csv)

REGION	SURROGATE	Cols 3-9	MERGE FUNCTION	SECONDARY SURROGATE	TERTIARY SURROGATE	QUARTER NARY SURROGATE	DETAILS	COMMENTS
USA	Population	…					Total population from Census 2000 blocks	
USA	Urban Population	…		Population			Total urban population from Census 2000	URBAN2 in Surrogate Source sheet.
USA	Residential Heating - Natural Gas	…		Housing			Number of Housing Units using Utility Gas for primary heating	
USA	Total Road Miles	…		Population			Sum of rural primary, urban primary, rural secondary and urban secondary road miles.	
USA	Urban Primary Road Miles	…		Total Road Miles	Population		Road Miles of Urban Primary Roads	No FRDCLASS. Should use NEWRD_CLAS

								and CLASS1_03_ ?
USA	0.75 Total Roadway Miles plus 0.25 Population	…	0.75*Total Road Miles+ 0.25*Population	Population			Combination of 3/4 total road miles surrogate ratio and 1/4 population surrogate ratio	
USA	Land	…					Land Area *data for this surrogate is contained in SMOKE-ready bgpro files, not ampro files.	
USA	Water	…		Navigable Waterway Activity	Navigable Waterway Miles	Land	Water area	
USA	Rural Land Area	…		Land			Land Area that is not within an area designated as an Urbanized Area or an Urban Cluster.	No rural_land in Sources of Surrogate sheet (it use us_urban).
USA	Total Agriculture	…		Rural Land Area	Land		Sum of: Pasture/Hay, Grains, Row Crops, Fallow Land and Orchards/Vineyards	
USA	Industrial Land	…		Urban Population	Land	Population	Sum of building square footage: IND1 + IND2 + IND3 + IND4 + IND5 + IND6	
USA	Heavy and High Tech Industrial (IND1 + IND5)	…		Industrial Land	Urban Population	Population	Sum of building square footage from FEMA categories: IND1 + IND5	Total Industrial in Table1. Same as Industrial Land?
USA	Forest External	…	0.5*../output/ US36KM_20X20/forest.txt| Forest External+ 0.5*Rural Land Area	../output/ US36KM_20X20/ mypop_100.txt| My Population				
NA	Population	…	Population[USA]; Population[Canada];					

			Population[Mexico]					

Table 5. Example of a Surrogate Generation Control File Loaded into a Spreadsheet

REGION	SURROGATE	SURROGATE CODE	GENERATE	QUALITY ASSURANCE
USA	Population	100	YES	YES
USA	Urban Population	120	NO	YES
USA	Residential Heating - Natural Gas	150	NO	YES
USA	Total Road Miles	240	NO	YES
USA	Urban Primary Road Miles	200	NO	YES
USA	0.75 Total Roadway Miles plus 0.25 Population	255	YES	NO
USA	Land	340	NO	YES
USA	Water	350	NO	NO
USA	Rural Land Area	400	YES	YES
USA	Total Agriculture	310	YES	YES
USA	Industrial Land	505	NO	YES
USA	Heavy and High Tech Industrial (IND1 + IND5)	570	NO	YES
USA	Forest external	328	YES	NO
NA	Population	100	YES	NO

4. Running the Surrogate Tool

4.1 Surrogate Tool

You can specify all input and control information for the Surrogate Tool easily using text editors or spreadsheet software. The Surrogate Tool runs on any operating system that supports Java and can run the Spatial Allocator. It has been tested on Linux. In order to run the Surrogate Tool, you must have the Java 2 Platform Standard Edition 5.0 or higher installed on your computer. If this is not already available, it can be downloaded from Sun's web site at: http://java.sun.com/javase/downloads/index.jsp

Once Java is installed, the Surrogate Tool can be started using a single command line argument—the location of the global control variables file, as shown in the following example:

```
java -classpath SurrogateTools.jar
gov.epa.surrogate.SurrogateTool control_variables_grid.csv
```

The Surrogate Tool reads the input files and then calls the surrogate creation, merging, and gapfilling programs as needed to generate each surrogate. The Tool verifies that the input files have the correct syntax. Note that you do not have to edit any scripts during this process, nor do you need to know the detailed requirements regarding the GIS functions involved.

The Surrogate Tool attempts to generate all surrogates for which appropriate input data are provided. If there are errors in the input specification for a particular surrogate, that surrogate is not generated during the run, but the Surrogate Tool continues to try to generate the remaining surrogates. The surrogate files are placed in the OUTPUT DIRECTORY you specify in the global control variables file. It is recommended that the grid name be included in the name of the output directory for regular grid or E-Grid surrogates. As the surrogates are created, quality

assurance information (e.g., surrogate numerators and denominators) is added to the surrogate files, if this has been requested in the generation control file with the QUALITY ASSURANCE variable. Comment lines that describe the newly created surrogates are also included in the file.

Each spatial surrogate is output to a separate surrogate file in the specified output directory. Appropriate SMOKE-required header information for the surrogate (e.g., #GRID or #POLYGON) is placed in each output surrogate file. The individual surrogate files that are produced by the tool are named according to the convention:

Region_code_NOFILL.txt (for non-gap-filled surrogates), or
Region_code_FILL.txt (for gap-filled surrogates)

The Surrogate Tool creates a log file that contains a summary of all the surrogates that were created at the bottom of it. If the creation of some surrogates failed, the execution of the Surrogate Tool can be restarted by providing an updated generation control file with GENERATE for only the unfinished surrogates set to YES.

Some intermediate text files are generated during the course of a run of the surrogate tool. They are placed in a subdirectory of the OUTPUT_DIRECTORY called "temp_files". The Surrogate Tool automatically creates this subdirectory. It is recommended that you keep these files because they are a record of scripts to run and all the files input to srgcreate and the merging, and gapfilling programs during the course of the run. You may also find these helpful for "debugging" purposes if things do not look right for one of the surrogates. Any old intermediate files will automatically be overwritten with the latest data during successive runs of the tool written to the same OUTPUT_DIRECTORY and are kept separate for each surrogate and region combinations, so you do not need to worry about deleting files between runs. The OVERWRITE OUTPUT FILES variable in the global control variables file does not control whether the files under the temp_files directory are overwritten.

4.2 Normalization Tool

This program that "normalizes" surrogates for counties that do not sum to 1 and makes the sum approximately 1. This should be used with care because surrogate values for counties / regions on the edge of the grid often should not sum to 1. The tool accepts an exclude list of such counties. Run the normalization tool with one of these commands:

```
java -classpath SurrogateTools.jar
gov.epa.surrogate.normalize.Main ../output/somegrid/SRGDESC.txt
exclude_list tolerance

java -classpath SurrogateTools.jar
gov.epa.surrogate.normalize.Main ../output/somegrid/SRGDESC.txt
```

[the default tolerance if left unspecified is 1e-6]

5. Output Files

The spatial surrogate files output from the Surrogate Tool contain the spatial allocation factors for nonpoint/area sources and non-link mobile sources. The surrogate files are ready to be used in SMOKE as AGPRO or MGPRO files, which are now read by SMOKE from the SRGDESC file. There are two output formats for computed surrogate ratios: one for grids (used for both Regular Grid and EGrid formats) and the other for polygon-based data such as census tracts. The format of the output surrogate file for regular grid surrogates is described in Table 6, and an example is provided in the Table 7. External surrogates input to the tool are also assumed to be in this format. At the time that this document was written, SMOKE does not support polygon surrogates.

In the surrogate file, the header line that describes the grid is followed by lines that describe how the surrogate in the file was computed, and the lines containing the surrogate fractions follow the comment lines. The numerator, denominator, and QA sum at the end of each line are optionally output by srgcreate when QUALITY ASSURANCE is set to YES for the surrogate. These values are preceded by a '!' to indicate that they are comments and are ignored by SMOKE. The numerator and denominator are values used to compute the surrogate fraction, and the QA sum is a running sum of the fractions for a given county. Typically this should be 1 for the last entry (e.g., the last grid cell or polygon listed) for a given county. The output file format for polygon-based surrogates is shown in Table 8, followed by an example in Table 9.

The surrogate files output from the srgcreate and merge tool programs are named according to the format: *region_code*_NOFILL.txt. If a surrogate is to be gapfilled, the gapfilled surrogate file will be created and named *region_code*_FILL.txt. The NOFILL files are not deleted because they are used as inputs for gapfilling or merging with other surrogates and they are useful for quality assurance purposes.

Several other types of output files are also created by the surrogate tool:

5.1 Surrogate Description File

A **Surrogate Description file**, which specifies the region, name, code, and final (i.e., after merging and gapfilling) file names of the individual spatial surrogate files created by the tool. This file is known to SMOKE as the SRGDESC file. If a surrogate was not gapfilled, this file contains the name of the NOFILL surrogate file for that surrogate ID, otherwise it contains the name of the FILL surrogate file. This is illustrated in the example of this file that is given in Table 10.

5.2 Log File

A **log file** that contains all information written by the tool itself, all of the output and error information produced by the Spatial Allocator and the gapfilling and merging programs, along with a summary of the generation of each surrogate. A summary of the surrogate computation with the regions, names, and codes are output to the end of the log file. So, users should check the end of the log file first to see the status of all surrogate computation. If some surrogate

computations fail, users can check the detailed log information above. An example is given in Table 11.

5.3 Output Surrogate File

If requested by the **OUTPUT SURROGATE FILE** keyword in the global control variables file, a file containing all surrogates is created by concatenating all the individual surrogate files included in the SRGDESC file. SMOKE versions 2.3 and higher do not require surrogates to be found in the same file, but older versions do. The headers for the concatenated surrogates are mixed in with the file; they are not all placed at the top of the file. Also, if you are using an older version of SMOKE prior to 2.2, the additional comment lines in the middle of the file will probably need to be removed.

1. All **intermediate text files** used as input to srgcreate tool are stored in the temp_files subdirectory of the OUTPUT_DIRECTORY. It is a good idea to keep these files around for debugging purposes and as a record of how the surrogates were created by srgcreate tool.

2. **Script files** (.csh for Linux system) for each surrogate computation using srgcreate are also created and stored in the same directory. Users can optionally use these scripts to run the Spatial Allocator in "standalone" mode, or to verify how the surrogate is computed by examining the values of the environment variables.

3. A shapefile containing the sum of the surrogate numerators for each grid cell or polygon is output to a file named grid *region_code, e*grid *region_code* or poly *region_code* for each surrogate computed from srgcreate. Essentially this file contains a gridded version of your surrogate weight data (e.g. gridded population). A corresponding CSV file of the attribute data is also created.

Table 6. Format of a Regular Grid Output Surrogate File

Line	Column	Description
1	A	#GRID
	B	Grid name
	C	X origin in units of the projection
	D	Y origin in units of the projection
	E	X direction cell length in units of the projection
	F	Y direction cell length in units of the projection
	G	Number of columns
	H	Number of rows
	I	Number of boundary cells
	J	Projection types (LAT-LON or LATGRD3, LAMBERT or LAMGRD3, UTM or UTMGRD3)
	K	Projection units
	L	Projection alpha value
	M	Projection beta value
	N	Projection gamma value

	O	X-dir projection center in units of the projection
	P	Y-dir projection center in units of the projection
2	A	#SRGDESC=
	B	Surrogate code
	C	Surrogate name
Remaining comment lines	A	#[Surrogate Generation Variable] =
	B	Value
Remaining lines	A	Spatial Surrogate code
	B	Country/State/County Code (Text or Integer)
	C	Grid column number (Integer)
	D	Grid row number (Integer)
	E	Spatial surrogate ratio

Table 7. A Sample Output Regular Grid Spatial Surrogate File

```
#GRID   US36KM_148X112  -2736000.000000 -2088000.000000 36000.000000    36000.000000
148     112     1       LAMBERT meters  33.0
00000   45.000000       -97.000000      -97.000000      40.000000
#SRGDESC=120,Urban Population
#
#SURROGATE REGION = USA
#SURROGATE CODE = 120
#SURROGATE NAME = Urban Population
#DATA SHAPEFILE = county_pophu2k
#DATA ATTRIBUTE = FIPSSTCO
#WEIGHT SHAPEFILE = pophu2k
#WEIGHT ATTRIBUTE = URBAN
#WEIGHT FUNCTION =
#FILTER FUNCTION =
#
#CONTROL VARIABLE FILE = /srgtool/control_variables.csv
#SURROGATE SPECIFICATION FILE = /srgtool/surrogate_specification.csv
#SHAPEFILE CATALOG = /srgtool/shapefile_catalog.csv
#GENERATION CONTROL FILE = /srgtool/surrogate_generation.csv
#SURROGATE CODE FILE = /srgtool/surrogate_IDs.txt
#GRIDDESC = /srgtool/GRIDDESC.txt
#
#USER = lran
#COMPUTER SYSTEM = linux
#DATE = Tue Sep 20 20:14:26 EDT 2005
# THE FOLLOWING LINE IS NOT PART OF THE ACTUAL OUTPUT BUT WAS ADDED FOR EXPLANATION
# SRGID  FIPS      COL    ROW     FRAC                NUMERATOR DENOMINATOR  QASUM
  120   53073      25      92   0.000752897     !     85.0819   113006       0.0007529
  120   53073      24      93    0.0142783      !     1613.53   113006       0.015031
  120   53073      25      93     0.927497      !      104813   113006       0.94253
  120   53073      24      94    0.0442883      !     5004.85   113006       0.98682
  120   53073      25      94    0.0131839      !     1489.86   113006       1
  120   53009      20      91   0.00927792      !     312.768   33711       0.0092779
  120   53009      21      91   0.00159502      !     53.7697   33711       0.010873
  120   53009      22      91     0.384065      !     12947.2   33711       0.39494
  120   53009      23      91     0.274769      !     9262.75   33711       0.66971
# DENOMINATOR_THRESHOLD CAME INTO PLAY IN THE FOLLOWING LINE
# 120   01075      99      40     0.419329      !     2.342e-7   5.587e-7   0.419329
```

Table 8. Format of a Polygon Surrogate File

Line	Columns	Description
1	A	#POLYGONS
	C	OUTPUT_POLY_FILE
	D	OUTPUT_POLY_ATTR
	E	OUTPUT_FILE_ELLIPSOID
	F	OUTPUT_FILE_MAP_PRJN
2	A	#SRGDESC=
	B	Integer Surrogate code
	C	Surrogate name

Remaining comment lines	A	#[Surrogate Generation Variable] =
	B	Value
Remaining lines	A	Spatial surrogate code (Integer)
	B	Country/State/County Code (Text or Integer)
	C	Unique polygon (e.g., Census Tract) ID (Text or Integer)
	D	Spatial surrogate decimal fraction (i.e., fraction of the surrogate attribute in the polygon) (Real)

Table 9. A Sample Output Census Tract (Polygon) Surrogate File

```
#POLYGON        OUTPUT_POLY_FILE=/emiss_shp2003/us/tnnc
OUTPUT_POLY_ATTR=STFID  OUTPUT_FILE_ELLIPSOID=SPHERE
OUTPUT_FILE_MAP_PRJN=+proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97
#SRGDESC=100,Population
#
#SURROGATE REGION = USA
#SURROGATE CODE = 100
#SURROGATE NAME = Population
#DATA SHAPEFILE = county_bndy
#DATA ATTRIBUTE = FIPSSTCO
#WEIGHT SHAPEFILE = pophu2k
#WEIGHT ATTRIBUTE = POP2000
#WEIGHT FUNCTION =
#FILTER FUNCTION =
#
#CONTROL VARIABLE FILE = ./control_variables_poly.csv
#SURROGATE SPECIFICATION FILE = ./surrogate_specification.csv
#SHAPEFILE CATALOG = ./shapefile_catalog.csv
#GENERATION CONTROL FILE = ./surrogate_generation.csv
#SURROGATE CODE FILE = ./surrogate_codes.csv
#
#USER = lran
#COMPUTER SYSTEM = linux
#DATE = Wed Nov 16 14:02:09 EST 2005
  100   51810   37053110101     1.22752e-22  !    5.22011e-17  425257       1.2275e-22
  100   51810   37053110102     4.44544e-20  !    1.89045e-14  425257       4.4577e-20
  100   51800   37029950100     4.06554e-19  !    2.58881e-14   63677       4.0655e-19
  100   51800   37073970200     5.21778e-18  !    3.32252e-13   63677       5.6243e-18
  100   51800   37073970300     1.65589e-18  !    1.05442e-13   63677       7.2802e-18
  100   51175   37073970300     5.17682e-26  !    9.05012e-22   17482       5.1768e-26
  100   51175   37091950100     6.57034e-18  !    1.14863e-13   17482       6.5703e-18
  100   51175   37131980100     3.73904e-18  !    6.53659e-14   17482       1.0309e-17
  100   51550   37053110200     1.54769e-18  !    3.08275e-13  199184       1.5477e-18
  100   51550   37029950100      3.729e-19   !    7.42757e-14  199184       1.9206e-18
  100   51025   37185950100     2.93599e-18  !    5.40781e-14   18419       2.936e-18
  100   51081   37131980300     1.67322e-17  !    1.93424e-13   11560       1.6732e-17
```

Table 10. An Example SRGDESC FILE for a RegularGrid*

```
#GRID   US36KM_148X112  -2736000.000000 -2088000.000000 36000.000000   36000.000000   148
112   1     LAMBERT meters  33.000000     45.000000     -97.000000     -97.000000     40.000000
USA,100,"Population",/output/US36KM_148X112/USA_100_NOFILL.txt
USA,120,"Urban Population",/output/US36KM_148X112/USA_120_FILL.txt
USA,130,"Rural Population",/output/US36KM_148X112/USA_130_FILL.txt
USA,137,"Housing Change",/output/US36KM_148X112/USA_137_NOFILL.txt
USA,140,"Housing Change and Population",/output/US36KM_148X112/USA_140_NOFILL.txt
USA,150,"Residential Heating - Natural Gas",/output/US36KM_148X112/USA_150_FILL.txt
USA,160,"Residential Heating - Wood",/output/US36KM_148X112/USA_160_FILL.txt
USA,170,"Residential Heating - Distillate Oil",/output/US36KM_148X112/USA_170_FILL.txt
USA,180,"Residential Heating - Coal",/output/US36KM_148X112/USA_180_FILL.txt
USA,190,"Residential Heating - LP Gas",/output/US36KM_148X112/USA_190_NOFILL.txt
USA,200,"Urban Primary Road Miles",/output/US36KM_148X112/USA_200_FILL.txt
USA,210,"Rural Primary Road Miles",/output/US36KM_148X112/USA_210_FILL.txt
USA,220,"Urban Secondary Road Miles",/output/US36KM_148X112/USA_220_FILL.txt
USA,230,"Rural Secondary Road Miles",/output/US36KM_148X112/USA_230_FILL.txt
```

* header line has been wrapped to two lines for this example

Table 11. A Sample Log File Created by the Surrogate Tool for RegularGrid Output

```
Run Date: Thu Mar 05 16:25:26 EST 2009
            Main Control CSV File
GENERATION CONTROL FILE ./surrogate_generation_grid.csv
SURROGATE SPECIFICATION FILE    ./surrogate_specification_2002.csv
SHAPEFILE CATALOG       ./shapefile_catalog.csv
SHAPEFILE DIRECTORY     ../data/emiss_shp2003/us
SURROGATE CODE FILE     ./surrogate_codes.csv
SRGCREATE EXECUTABLE    ../bin/srgcreate.exe
DEBUG_OUTPUT    Y
OUTPUT_FORMAT   SMOKE
OUTPUT_FILE_TYPE        RegularGrid
OUTPUT_GRID_NAME        M08_NASH
GRIDDESC        ./GRIDDESC.txt
OUTPUT_FILE_ELLIPSOID   +a=6370000.0,+b=6370000.0
OUTPUT DIRECTORY        ../output/somegrid
OUTPUT SURROGATE FILE   ../output/somegrid/allsrgs.txt
OUTPUT SRGDESC FILE     ../output/somegrid/SRGDESC.txt
OVERWRITE OUTPUT FILES  YES
LOG FILE NAME   srg_grid.log
DENOMINATOR_THRESHOLD   0.0005
COMPUTE SURROGATES FROM SHAPEFILES      YES
MERGE SURROGATES        YES
GAPFILL SURROGATES      YES

            Get Grid Header For Surrogate Files
SRGCREATE_ERROR>WARNING: Environment variable: MAX_LINE_SEG, not set
SRGCREATE_OUTPUT>MIMS Surrogate Creator Version 3.5, 8/12/2008
SRGCREATE_OUTPUT>
```

```
SRGCREATE_OUTPUT>EV: OUTPUT_FILE_TYPE=RegularGrid
SRGCREATE_OUTPUT>Setting output grid
SRGCREATE_OUTPUT>
SRGCREATE_OUTPUT>EV: OUTPUT_FILE_TYPE=RegularGrid
SRGCREATE_OUTPUT>Reading Regular Grid
SRGCREATE_OUTPUT>
SRGCREATE_OUTPUT>EV: OUTPUT_GRID_NAME=M08_NASH
SRGCREATE_OUTPUT>MAX_LINE_SEG not set, discretization intervals disabled
SRGCREATE_OUTPUT>griddesc file name = ./GRIDDESC.txt
SRGCREATE_OUTPUT>
SRGCREATE_OUTPUT>Ellipsoid var = OUTPUT_FILE_ELLIPSOID
SRGCREATE_OUTPUT>EV: OUTPUT_FILE_ELLIPSOID=+a=6370000.0,+b=6370000.0
SRGCREATE_OUTPUT>Ellipsoid=+a=6370000.0,+b=6370000.0
SRGCREATE_OUTPUT>EV: OUTPUT_GRID_NAME=M08_NASH
SRGCREATE_OUTPUT>Not using BB optimization
SRGCREATE_OUTPUT>
SRGCREATE_OUTPUT>EV: OUTPUT_FILE_TYPE=RegularGrid
SRGCREATE_OUTPUT>#GRID  M08_NASH       1000000.000000  -536000.000000  8000.000000
8000.000000    46     42     1      LAMBERT meters  30.000000       60.000000       -100.000000     -
100.000000     40.000000
SUCCESS IN RUNNING THE EXECUTABLE: SRGCREATE

End Date: Thu Mar 05 16:25:27 EST 2009
Elapsed time in minutes: 0.008583333333333333

SUCCESS -- The Program Run Completed
```

6. Development Description

6.1 Integration with the Emissions Modeling Framework

The following is a summary of the features of the Surrogate Tool and of its integration with the Emissions Modeling Framework (EMF):

1. Shapefiles are sources of geographic data used to create spatial surrogates. Users will be able to view, add to, and remove from a list of available shapefiles via the EMF's data management capabilities. For the stand-alone tool, an ASCII file serves as a catalog for shapefiles accessible on local disks.
2. Users can define the spatial surrogates to create, and the shapefiles used to create them, in the CSV format configuration files for the Surrogate Tool. These files can easily be edited using spreadsheet software such as Excel, and these input files can be loaded into the EMF data management system.
3. The Surrogate Tool can generate spatial surrogates using shapefiles or spatial surrogates generated internally or externally. The tool outputs the surrogates in formats used by SMOKE for grid-based or polygon-based modeling. In a single run of the tool, users can make surrogates either for a regular grid, E-Grid or for polygons. The output surrogates are self-describing, so the origin of the surrogate ratios is discernible. The header and

comment lines include enough information to allow you to regenerate the surrogates. The comment lines include which shapefiles were used, any filter or weight functions that were applied, attributes that were used, merge function used, etc.

4. The EMF can import the surrogate files created by the Surrogate Tool the SRGDESC file, and the log file. The EMF can also export all of these files prior to running SMOKE.
5. Once the EMF can support executing external programs, users will be able to execute the Surrogate Tool from the EMF, and the resulting output surrogate file(s) will be automatically imported into the EMF as datasets.
6. The Surrogate Tool can combine existing surrogates generated using other packages with surrogates generated by the tool if the header line with #GRID or #POLYGON are the same. These externally created surrogates can be used as input to merging or gap filling, or concatenated into a merged surrogate file. The resulting file works in the same way as if all surrogates were generated with the tool. For example, if you have two surrogate files (both containing the same surrogate) that were already created, then the tool can place one or both of those in a new file that also includes surrogates generated by the tool. You may be required to add or modify the header line with #GRID or #POLYGON (such as grid name and projection information) to the externally generated surrogate files prior to using them within the system. The Surrogate Tool verifies that the external files are based on the same grid or polygon for which the Surrogate Tool is being run. The tool does not require all surrogates in the externally generated surrogate file to be used, but extracts all surrogate fractions for a specified code.
7. The outputs from srgcreate and the merging and gapfiling tools, along with some additional summary information, are placed in a log file created by the Surrogate Tool.
8. The EMF can import and export a gridding cross reference that can be used by SMOKE.
9. Through the EMF, users can update gridding cross-reference data to use newly created surrogates, by indicating which inventory characteristics (e.g., SCC) map to the new surrogate.

6.2 Program Logic

The Surrogate Tool software takes the following steps when it runs:

1. Read in the global control variables file and store the global variables and their values in memory. The global control variables file specifies the names of the four other CSV input files and one text file to read, along with other information regarding the generation of surrogates.
2. Read the specified surrogate specifications file (SFF) and store the surrogate definitions into memory.
3. Read the specified shapefile catalog file and store all shapefile descriptions into memory.
4. Read the specified generation control file and store the surrogates to generate into memory.

5. Read the specified surrogate code file and store the surrogate name and code information into memory.
6. Check the contents of these files as follows:
    a. Ensure that the output directories exist; otherwise the tool will create the new directory specified.
    b. Ensure that the shapefiles in the SSF match the shapefile names in the catalog.
    c. Ensure that all files needed exist.
    d. Ensure that there are no redundant entries in the input files
7. Get main environment variables, which are the same for each surrogate computation (e.g. grid, output directory).
8. Obtain the header with #GRID or #POLYGON which is the same for each surrogate.
9. If COMPUTE SURROGATES FROM SHAPEFILES variable in the global control variables file is set to YES, the program will loop through the generation control file to compute surrogate ratios from shapefiles using srgcreate for surrogates with GENERATE set to YES. Any surrogates for which the value in the GENERATE column is not YES will not be computed.
10. For each surrogate that depends on shapefiles, the following steps are taken:
    a. Shapefile information is looked up in the shapefile catalog. If the definition of the shapefile is not found, an error is issued and the program moves on to generating the next surrogate.
    b. The tool obtains all needed environment variables to run srgcreate. If the surrogate computation uses a filter function (e.g., include only shapes for which the ROAD_TYPE=4), a filter text file that can be used with srgcreate will be generated. All filter text files are stored in the temp_files subdirectory of the OUTPUT_DIRECTORY and are named *filter_region_code.txt.*
    c. The tool runs srgcreate with all environment variables for the surrogate to be computed.
    d. The processing information by srgcreate written to standard output and standard error is copied into the log file.
    e. The tool checks whether the computation is successful. If the computation is successful, the header and comment information will be inserted to the beginning of the output file. The output file for this surrogate is saved in the OUTPUT DIRECTORY defined in the global control variables file using a predefined file name, *region_code*_NOFILL.txt. If the computation fails, an error message is output to the log file and the program moves on to the next surrogate computation.
    f. srgcreate outputs a grid or poly output shapefile with the sum of the surrogate numerators using pre-defined name – *grid_region_code* for regular grid output, egrid_*region_code* for egrid output or *poly_region_code* grid for polygon output in the OUTPUT DIRECTORY. A CSV file with grid ID (column and row) or polygon ID (such as census tract ID) and surrogate ratio is also created and stored in the same

directory. The CSV file uses a pre-defined name *grid_region_code.csv, egrid_region_code.csv or poly_region_code.csv* for output.

g. A script that can be used to regenerate the surrogate is created and named either *region_code*_NOFILL.bat or *region_code*_NOFILL.csh in the temp_files subdirectory.

11. Once all surrogates based on weight shapefiles have been created, if the MERGE SURROGATES variable in the global control variables file is set to YES, the program loops through the generation control file again to find any surrogates with GENERATE set to YES that use a merge function in the surrogate specification file.

12. For each surrogate to be created based on a merge function, the following steps are taken:

    a. The tool obtains all needed environment variables to run the merging tool. A merge text file to be used will be generated. The merge text files will be stored in the temp_files subdirectory of the OUTPUT_DIRECTORY and they are named merge_*region_code*_NOFILL.*txt.*

    b. All needed surrogate files are checked for existence. If any error occurs, the program will move to next surrogate merging.

    c. The tool runs the merging tool with the merge text file and all other environment variables for the surrogate to be computed.

    d. The processing information by the merging tool written to standard output and standard error is copied into the log file.

    e. The tool checks whether the computation is successful. If the computation is successful, the header and comment information will be inserted to the beginning of the output file. The output file for this surrogate is saved in the OUTPUT DIRECTORY defined using a predefined file name, *region_code*_NOFILL.txt. If the computation fails, an error message will be output to the log file and the program will move to the next surrogate computation.

13. Once all surrogates that depend on shapefiles and merge functions have been created, the Surrogate Tool performs gap filling on those surrogates using the gapfilling tool. If GAPFILL SURROGATES variable in the global control variables file is set to YES, the program loops through the generation control file again to find any surrogates with GENERATE set to YES for which there are secondary, tertiary, or quarternary surrogates in the surrogate specification file – these are the surrogates that must be gapfilled.

14. For each surrogate to be gapfilled, the following steps are taken:

    a. The tool obtains all needed environment variables to run the gapfilling tool. A gapfill text file to be used is generated. The gapfill text file is stored in the temp_files subdirectory of the OUTPUT_DIRECTORY and is named gapfill_*region_code.txt.*

    b. All needed surrogate files are checked for existence. If any error occurs, the program moves to the next surrogate to be merged.

c. The tool runs the gapfilling tool using the gapfill text file and all other environment variables required for gapfilling.

d. The information output by the gapfilling tool is written to standard output and standard error is copied into the log file.

e. The tool checks whether the computation is successful. If the computation is successful, the header and comment information will be inserted to the beginning of the output file. The output file for this surrogate is saved in the OUTPUT DIRECTORY using a predefined file name, *region_code*_FILL.txt. If the computation fails, an error message will be output to the log file and the program will move to the next surrogate gapfilling.

15. The Surrogate Tool keeps information about how the surrogate is computed. After each surrogate is computed, merged, or gapfilled, the program will create or update the SRGDESC file defined in the global control variable CSV file. The SRGDESC file will be used to tell SMOKE the location of all of the surrogates. For example, if the surrogate is already in the SRGDESC file, the current computed file will replaced the old file. If the surrogate is not in the SRGDESC file, the surrogate with the computed file will be add to the SRGDESC file. If a surrogate is gapfilled after being computed from a surrogate shapefile or from a merge function, only the gapfilled surrogate file is listed in the SRGDESC file.
16. Information is written to a log file, which includes a summary table of the surrogate computation that is written to the bottom of the log file (see Table 11). This lists the region, name, and code for each of the surrogates that were requested to be generated. It also indicates whether the computation of srgcreate, merging or gapfilling was successful or failed for each surrogate.
17. If the value of the OUTPUT SURROGATE FILE variable specified in the global control variable file is not NONE, a concatenated file of all generated surrogates from the SRGDESC file is created.
18. If any error occurs in the program run, the final exit status of the Surrogate Tool is nonzero. If all of the requested surrogates were created successfully, the exit status is zero.

7. Enhancements, Limitations and Future Updates

The following enhancements to the Spatial Allocator were made as part of the development of the Surrogate Tool.

1. Srgcreate, merging, and gapfilling were updated and enhanced to handle surrogate ratio computation for polygon output modeling (such as census tracts) and egrid WRF/NMM-CMAQ modeling in addition to the standard regular grid output.

2. Srgcreate was updated to handle multiple shape entries with the same attribute ID when reading the base data polygon into memory.

3. Srgcreate was updated to handle projection comparison and geographic ellipsoid comparison.

4. Srgcreate was modified to handle the new environment variable: DENOMINATOR_THRESHOLD for surrogate computation.

5. Srgcreate was updated to output surrogates as comments when there was no data polygon ID.

6. The merging tool was updated to output surrogates with 8 decimal places instead of 6.

7. The merging and gapfilling tools were re-developed in Java and they are packed as part of the SurrogateTools jar file. There is no assumption regarding the complete list of counties being found in the lowest level surrogate in the new gapfilling tool as there is in srgmerge version of gapfilling. The Java-based merging and gapfilling tools were improved to work for polygon output modeling (such as census tracts).

8. A Java quality assurance and summary reporting tool was developed. The tool summarizes all surrogates listed in SRGDESC.txt file based on surrogate codes and counties. For each region and county in the surrogate files it reports on how the surrogates were gapfilled, surrogates that do not sum to 1, and counties for which there is no data available.

The Surrogate Tool has some limitations. The following list could be future enhancements or updates for improving the Surrogate Tool:

1. Update the merging and gapfilling tools to accept surrogate codes in their input files so that the "surrogate code CSV file" does not need to be provided as an additional input file to the surrogate tool.

2. Support use of data from state-specific shapefiles. This would make it easier to make updates of the data by state. Additionally, we could support the optional splitting of files into state specific ones as a preprocessing step using a State shapefile and/or attributes like FIPS_CODE. Users may wish to run for a subset of the states or all states that overlap the grid, which would make the runs more efficient.

3. Currently, the shapefile catalog CSV file contains projection information for each shapefile. We may wish the tools read .prj components from the shapefiles. So, the projection information will be read from .prj file of the shapefile instead of being specified by the users. The requirement is that all shapefiles used have to be well documented with projection information. Some of surrogate shapefiles we downloaded from the EPA web sites do not have .prj files. If the Spatial Allocator can read .prj files, it will make easier to specify and track the map projections of shapefiles in the computation. Currently, the PROJ.4 syntax is used to specify map projection and ellipsoid information in the shapefile catalog.

4. A function for geographic transformation in srgcreate could be added to handle two different geographic datum systems. Right now, the ellipsoid definition of base and weight shapefile is compared with the output ellipsoid. If they are different, the program will exit with an error. Users have to transform the ellipsoid of the base or weight

shapefile into the output ellipsoid externally in order to correctly project coordinates of base or weight shapefile.