New Tools to Generate Spatial Surrogate and Speciation Profile Inputs to SMOKE

Alison M. Eyth, Limei Ran, Rajasooriyar Partheepan
Carolina Environmental Program, University of North Carolina at Chapel Hill
137 E. Franklin St., Chapel Hill, NC 27599-6116
eyth@unc.edu

Greg Yarwood, Michele Jimenez, Sunja Rao
ENVIRON International Corporation
101 Rowland Way, Suite 200, Novato, CA 94945
GYarwood@environcorp.com

ABSTRACT

As part of the development of EPA's new Emissions Modeling Framework (EMF), two new tools have been developed to aid emission modelers in the development of SMOKE input files for spatial surrogates and speciation profiles. The Surrogate Tool can create dozens of spatial surrogates for multiple regions without requiring the user to do any scripting. The Surrogate Tool is controlled using comma separated value (CSV) files that describe the surrogates to be created, while Shapefiles provide the input data. Support for the creation of polygon-based surrogates (e.g., for census tracts) was also added and will be used to interface with AERMOD. The Speciation Modeling Data Tool was developed to provide a way to generate speciation profiles for any defined chemical mechanism, to support both integrated and non-integrated HAPS inventories, and to provide the ability to define active and tracer toxics in the modeling inventories. The tool reads VOC and simplified PM2.5 profile weight information and species data exported from the SPECIATE 4.0 database, pre-computed speciation profiles for other pollutants, and chemical mechanism definitions as inputs. A model run consists of a keyword based control file that specifies run parameters. In addition, the user specifies whether or not the CAPS and HAPS inventories are integrated, and whether individual species are to be treated as active or tracer species. Given these inputs, the speciation tool computes the speciation factors for the specified chemical mechanism and will create the speciation profiles and VOC-to-TOG factors that will be used by SMOKE and the EMF.

INTRODUCTION

The purpose for developing the Emissions Modeling Framework is to solve many long-standing difficulties in emissions modeling at EPA. The goals of the framework are to (1) remove bottlenecks and prevent errors and reworking by EPA and its contractors, (2) create best practice approaches and protocols for emissions modeling, and (3) develop a long-needed software infrastructure for the EPA modeling community that will perform emissions modeling tasks in a consistent way across multiple disciplines, share emissions modeling data across EPA, and provide a transparent and self-documenting approach to emissions modeling. Additionally, it is hoped that the Framework will prove useful for modelers outside of EPA, where such benefits come naturally and without additional resource needs. The EMF will provide integrated quality control processes to foster high quality of emissions results, data handling, organization of data, tracking of emissions modeling efforts, and real-time accessibility of information. One goal of the EMF is to make it easier to produce, maintain, and track SMOKE ancillary files and their contents. The Surrogate Tool and Speciation Tool help to accomplish this goal.

The Surrogate Tool is a stand-alone 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 inputs to the Surrogate Tool are user-defined text files that specify the

surrogates to be generated and the methods to use to generate them. The formats of these input files allow them to be easily edited and maintained in a spreadsheet program like Microsoft Excel. These input files were purposefully designed to provide all relevant information needed to regenerate the surrogates, and therefore simultaneously provide documentation for all surrogates produced by the tool. This documentation does not need to be maintained separately from the input files to the tool, which eliminates the possibility of it becoming out of sync with the methods actually used to generate the surrogates.

The Surrogate Tool runs without a Graphical User Interface. It requires a single command line argument and can be run interactively or submitted to a queue to run in batch mode. The output from the Surrogate Tool is a log file, a set of ASCII SMOKE spatial surrogate ratio files that can be input to SMOKE for grid-based or polygon-based modeling, and a surrogate description (SRGDESC) file. Each surrogate created by the tool is placed into an individual file, and a concatenated file of all surrogates can also be created, if needed. Note that SMOKE versions 2.3 and higher support using spatial surrogates in separate files (one file per surrogate); while 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 and SMOKE uses this file to find all of the available surrogates. 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 Speciation Modeling Data Tool is a stand-alone tool for generating chemical speciation profiles that are inputs to air quality emission models such as SMOKE. The inputs to the Speciation Modeling Data Tool are text files; standard default files that define the chemical mechanisms, species properties, and weight profiles, as well as, a user generated control file and optional text files specifying lists of active or tracer species and integrated or non-integrated HAPS.

The Speciation Modeling Data Tool is a PostgreSQL database application. PostgreSQL is a powerful open sourced relational database management system that runs on all major operating systems; Linux, UNIX, and Windows. The Speciation Modeling Data Tool is initialized with a single command line script that creates a project database, loads the SQL modules, and imports a set of standard text files. Following initialization, a single command line Perl interface program is executed to run the application. The output from the Speciation Modeling Data Tool is the ASCII SMOKE speciation profiles file (GSPRO) and the VOC-to-TOG conversion factors (GSCNV) file. In addition, messages are displayed during execution to indicate progress through the computation steps.

SURROGATE TOOL

Background on the Spatial Allocator

The Surrogate Tool is built upon the Multimedia Integrated Modeling System (MIMS) Spatial Allocator (http://www.cep.unc.edu/empd/projects/mims/spatial/), which has the features needed to produce spatial surrogates, but can be complex to use. The MIMS Spatial Allocator programs provide an alternative to using commercial GISs or statistical software for generating spatial surrogates that are input to emissions models, merging surrogates, gapfilling surrogates, and performing other types of spatial allocation. A drawback of using commercial software to generate surrogates is that the software can cost thousands of dollars. In addition, these software systems are very complex and can be difficult to learn and use, and require custom coding to generate spatial surrogates. The MIMS Spatial Allocator programs are a comparatively small set of software written in C and C++, and can run on Windows, Linux, or other Unix systems. The programs were developed over several years and are run using operating specific scripts. The Spatial Allocator programs are provided free and use ESRI® Shapefiles—a standard in the GIS industry—as input databases. Spatial surrogates can be generated for point-, line-, or polygon-based data, sets such as ports, airports, housing, population, agriculture, water

area, and railroads. Additional information is provided in the Spatial Allocator documentation, available on the website listed above.

To create a surrogate, the Spatial Allocator program srgcreate uses a data file, a GRIDDESC file, and a weight file. The data file is a Shapefile that typically consists of county polygons. The GRIDDESC file contains descriptions of one or more air quality model grids, from which one is selected for a particular run of the Spatial Allocator. The weight file is a Shapefile that contains "shapes" that are points, lines, or polygons from which attributes are used (e.g. population) to weight the value of the surrogate in each of the grid cells of the air quality model such that the surrogate has larger values in cells for which the weight is highest. The values for a surrogate should sum to 1 for each county that is not on the edge of the modeling grid. The file produced by srgcreate contains a "surrogate" that is used by an emissions model (e.g., Sparse Matrix Operator Kernel Emission (SMOKE)) to allocate the emissions in a county into appropriate model grid cells. The mapping of the emissions to the surrogate is done using a cross reference file which specifies the appropriate surrogate to use for each type of emissions. For example, dry cleaning emissions are typically allocated using a drycleaner or population surrogate, whereas recreational boat emissions are allocated based on a water body surrogate. After processing by an emissions model, the resulting gridded, speciated, time-varying emissions are provided as an input to an air quality model such as the Community Multiscale Air Quality (CMAQ) model.

Aside from the srgcreate program which creates surrogates, there are several other Spatial Allocator programs that perform other functions. The beld3smk program can also process BELD3 data to create the inputs for biogenic processing that are required by SMOKE. The srgmerge program merges surrogates using functions (e.g., 0.25*Population + 0.75*Total road miles) and fills in missing data or "gaps" from surrogates with data from other surrogates. The allocator program performs general spatial allocation functions such as mapping gridded data to and from county-level data, mapping from grid to grid, and aggregating data from census tract to county levels. It can also convert Shapefiles from one map projection to another and print the attributes of the points, lines, or polygons that are overlapped by a grid, bounding box, or set of polygons.

Additions to the Spatial Allocator program were made in 2005 and early 2006. Srgcreate can now generate surrogates for polygons (e.g., census tracts), such as would be used by the ASPEN dispersion model. The Surrogate Tool interfaces with srgcreate and srgmerge to support the generation of both grid and polygon based spatial surrogates. New Java-based quality assurance, normalization, merging, and gapfilling tools are also available in the same package as the Surrogate Tool.

Need for the Surrogate Tool

The Surrogate Tool was developed to provide a more user-friendly way to use the Spatial Allocator which provides *built-in documentation of the methods used to compute the surrogates*. Because it has been developed using Java, the Surrogate Tool can generate surrogates regardless of the operating system on which it is used - as long as the Spatial Allocator programs can be compiled on that operating system. Unlike when using the Spatial Allocator stand-alone, users do not need to define UNIX environment variables, create intermediate text files, or use UNIX scripts to generate the surrogates. Instead, the Surrogate Tool uses Comma Separated Value (CSV) and other types of ASCII files to define how the surrogates should be computed, what input data should be used, where the results should be stored, which surrogates should be computed during the run, and other needed information.

Surrogate Tool Input Files

The Surrogate Tool's input files are five comma-separated-value (.CSV) files and a grid description (GRIDDESC) 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, or by a

standard text file editor. In addition, these files can be imported into the Emission Modeling Framework for data management and data versioning support. Examples of each of the input files are provided with the Surrogate Tool installation package and can be customized to meet your needs. 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 name of this file is given as the single command line argument to the Surrogate Tool.

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.

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.

4. 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).

5. 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.

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.

Examples of the input files as they would appear if loaded into a spreadsheet are shown in Tables 1-5. The surrogate specification file is the key input file because it controls the specifics how each surrogate is generated. In this file, the combination of REGION and SURROGATE must be unique. The DATA ATTRIBUTE is the attribute in the data (e.g., county) Shapefile that is used as the spatial identifier in the output surrogate file. Conceptually, the simplest surrogates to generate are those for which a WEIGHT ATTRIBUTE is specified. The weights for these surrogates are taken directly from a single column in the weight Shapefile. A value of NONE in this column tells the Spatial Allocator to use the area for polygons, length for lines, or point counts for points in the weight Shapefile instead of reading an attribute value. A WEIGHT FUNCTION may be used when the surrogate weights should be computed using a function of several attributes of the weight Shapefile. A FILTER FUNCTION is used when not all of the shapes in the Shapefile are to be included in the output surrogate. For example, if a Shapefile contains both land and water polygons and you wish to create a water surrogate, you could set a FILTER FUNCTION of H2O_CODE !=2, where the value 2 indicates a land polygon.

In some cases it is desirable to create a surrogate as a function of two other already computed surrogates. This is accomplished by specifying a MERGE FUNCTION (e.g., 0.5*Population + 0.5*Housing). Note that you may merge surrogates created outside of the surrogate tool by using the external file syntax of the form *surrogate_file* | *surrogate_name*. Note that the surrogate name to code mapping must be included in the surrogate code file. An example of this is the following:

0.5*../output/US36KM_20X20/forest.txt|Forest External+ 0.5*Rural Land Area

Merge functions should be limited to the form a*s1+b*s2 where a+b=1. Any surrogate (merged or computed using a Shapefile), may be gapfilled by specifying surrogate names or external files in the SECONDARY SURROGATE, TERTIARY SURROGATE, and QUARTERNARY SURROGATE columns. If data is missing from the primary computed or merged surrogate for a given county, values from the secondary surrogate will be used if available, else values from the tertiary surrogate or quarternary surrogate will be used.

Running the Surrogate Tool

The Surrogate Tool runs on any operating system that supports Java and can run the Spatial Allocator. It can be downloaded from http://www.cep.unc.edu/empd/projects/mims/spatial/srgtool. It has been tested Windows and Linux. The Spatial Allocator has also been run on Solaris, AIX, and IRIX. In order to run the Surrogate Tool, you must have the Java 2 Platform Standard Edition 5.0 installed on your computer. If this is not already available, it can be downloaded from Sun's web site at http://java.sun.com/j2se/1.5.0/download.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_file
```

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. 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 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.

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.

Output Files from the Surrogate Tool

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. An example of an output surrogate file for regular grid surrogates is shown in Table 6. External surrogates input to the tool are also assumed to be in the same format, excluding the comments. Note that versions of SMOKE 2.3 and earlier do not support polygon surrogates, but support for this is expected to be added in FY07.

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 county if the county is not on the edge of the modeling grid.

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:

1. 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 7.

2. 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 8.

3. 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.

4. All **intermediate text files** used as input to srgcreate and merging and gapfilling tools 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 and from the merging and gapfilling tools..

5. **Script files** (.csh for Linux system or .bat for window 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.

6. 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.

Other Related Tools

New Java-based quality assurance, normalization, merging, and gapfilling tools are available in the same package as the Surrogate Tool. The quality assurance tool provides tabular county level summaries of the sum of the surrogate fractions for each county, counties for which data is missing, and how each surrogate was gapfilled. Java implementations of the gapfilling and merging functions found in srgmerge are also provided. These Java implementations address some of the limitations and assumptions that were built into the srgmerge program. Space does not allow these tools to be discussed further here, but descriptions of them are available with the Surrogate Tool documentation at http://www.cep.unc.edu/empd/projects/mims/spatial/srgtool.

SPECIATION MODELING DATA TOOL

Background on the Tool

The Speciation Modeling Data Tool was developed to generate the split factors required for speciating emission inventory criteria pollutants to the lumped modeling compounds required for Air Quality Modeling. In particular, VOCs must be split into the lumped compounds of a particular chemical mechanism. The two most prevalent mechanisms in use are CBIV and SAPRC99. In the past, most Air Quality Modeling was performed based solely on criteria pollutants (CAPS). However, more and more frequently, the trend is to develop inventories that include both CAPS and HAPS. The Speciation Modeling Data Tool was designed specifically to support the availability of both CAPS and HAPS, to either integrate the HAPS in the chemical mechanism or not and to include both active and tracer species in modeling. The Speciation Modeling Data Tool allows for a new chemical mechanism definition (for example CB05) to be easily imported into the tool and then used to generate new speciation profiles based on this mechanism.

Speciation Modeling Data Tool

The Speciation Modeling Data Tool was developed to provide a way to generate speciation profiles for any defined chemical mechanism, to support both integrated and non-integrated HAPS inventories, and to provide the ability to define active and tracer toxics in the modeling inventories. Because it has

been developed using the open source PostgreSQL and Perl DBI, the Speciation Modeling Data Tool can generate profiles regardless of the operating system on which it is used - as long as PostgreSQL, Perl, and Perl DBI are installed. PostgreSQL can be downloaded from http://www.postgresql.org. Perl and the database drivers can be downloaded from http://www.perl.com. In order to run the Speciation Modeling Data Tool you must have database create privileges in PostgreSQL.

The Speciation Modeling Data Tool modules and metadata reside within the within the ***shared*** schema of the Speciation Modeling Data Tool PostgreSQL database. A schema is essentially a namespace: it contains named objects (tables, functions) whose names may duplicate other objects existing in other schemas. The *shared* metadata includes: mechanism definitions, profile descriptions, profile weights, and species data. The data flow of the Speciation Modeling Data Tool is displayed in Figure 1.

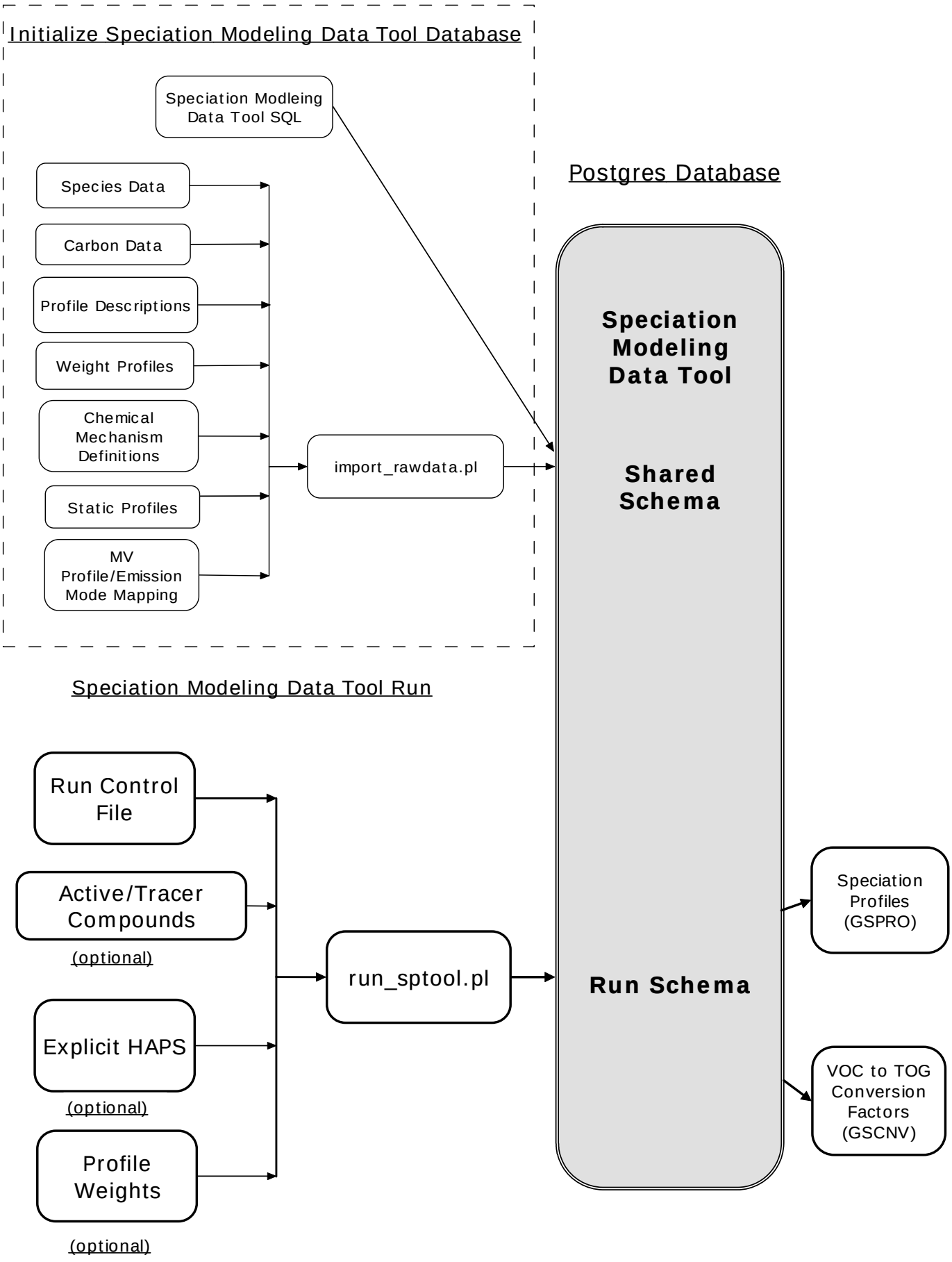
Speciation Modeling Data Tool Initialization

The ***init_db*** script file creates the Speciation Modeling Data Tool PostgreSQL database, loads the SQL functions, which are the work horse of the Speciation Modeling Data Tool, into the database, and imports the metadata. The *init_db* script can be edited to change the database name for a system installation, the default name is *sptoolv1_1*. *init_db* creates the *shared* schema which all Speciation Modeling Data Tool runs will access. The init_db script is displayed in Table 9.

The Speciation Modeling Data Tool *shared* schema consists of ten different data types. These default data are specified in text CSV formatted files and loaded into PostgreSQL through multiple calls to the Perl program, import_rawdata.pl. In order to optimize importing this data, if an error occurs while importing a file a database rollback is initiated and the import is aborted. The database is basically unrolled to where it was prior to the start of the import program.

1. The mechanism file provides records of mechanism name, species identifier, the corresponding lumped modeling compound identifier and the moles of model species per gram of emissions.
2. The carbon file provides the number of carbon bonds for each chemical mechanism lumped modeling compound.
3. The species file, exported from the SPECIATE 4.0 database, uniquely lists specie identifier, name, molecular weight, non-volatile weight and other historical data.
4. The gas profiles file, exported from the SPECIATE 4.0 database, uniquely lists gas profile identifier, name, description, documentation and numerous historical data.
5. The gas profile weights file, exported from the SPECIATE 4.0 database, lists gas profile identifier, specie identifier, the percent contribution, and some descriptive data.
6. The PM profiles file, exported from the SPECIATE 4.0 database, uniquely lists the simplified particulate profile identifier, name, description, documentation and numerous historical data.
7. The PM profile weights file, exported from the SPECIATE 4.0 database, lists simplified particulate profile identifier, specie identifier, the percent contribution, and some descriptive data.
8. The PM species file, lists the air quality model compound identifier for each specie identifier and emission inventory pollutant code for each air quality model. This is necessary as CAMx and CMAQ use different identifiers for PM components.
9. The static file, is the list of static profiles with corresponding split factor and divisor.
10. The process file, lists the on-road mobile emission modes by profile identifier. This supports the pollutant-emission mode naming convention used in SMOKE.

Figure 1. Speciation Modeling Data Tool Data Flow Diagram



Speciation Modeling Data Tool Input Files

The Speciation Modeling Data Tool's input files are a keyword-based control file and three optional comma delimited text files. Examples of the input files are shown in Tables 10 – 13. Descriptions of the input files are as follows:

1. The control file is a CSV file that specifies the model run information (e.g., modeling chemical mechanism, type of run, profile tolerance, output options, air quality model, optional input file names, and output file names). The name of this file is given as the single command line argument to the Speciation Modeling Data Tool.

2. The toxics file, an optional input, lists the toxic compounds that are explicit in the model's chemical mechanism and whether they are to be treated as active or tracer compounds. Active species are included in the chemical mechanism with chemical feedback, while tracer species are included in the mechanism with no chemical feedback. Tracer toxics species mass is double counted.

3. The HAPS file, an optional input, specifies the list of species identifiers and the corresponding emission inventory name for which HAPS inventory data will be included in the modeling.

4. The profiles file, an optional input, provides a list of new weight profiles for which to develop splits. If this file is specified in the run control file it overrides the use of the weight profiles in the *shared* schema.

Running the Speciation Modeling Data Tool

After the Speciation Modeling Data Tool is initialized a model run is made with a unique run identifier. The run identifier is used to create a run-based schema within the Speciation Modeling Data Tool database. This run identifier must begin with a character, must **not** be "*shared*", and should not be an already existing run identifier, unless an existing run is meant to be replaced. The run identifier becomes the schema name. All of the user specified parameters, run-specific inputs, and all of the temporary tables used to store intermediate calculations are stored in the run-based schema.

The *run_sptool.pl* program creates the run schema, reads and imports the user specified control file and run-specific data files, executes the SQL functions to compute the split factors and conversion factors, and writes the ASCII outputs. The command to run the Speciation Modeling Data Tool is:

```
perl run_sptool.pl database_name run_identifier run_control_file
```

Following is a methodology outline for the Speciation Modeling Data Tool. The basic calculations are the same for all run types. What makes a run type different is the way in which the chemical mechanism and the profile weights are initialized. The run control file contains a keyword RUN_TYPE which has the options of ALLCRITERIA, INTEGRATE or NOINTEGRATE. ALLCRITERIA means that all VOC model species, including toxic VOC species, are computed from criteria VOC emissions. INTEGRATE means that specified HAPS species are to be integrated from a separate inventory. Part of the criteria VOC mass is replaced with the HAP VOC mass (from the HAP part of the emission inventory). SMOKE computes the NONHAPVOC mass from the criteria VOC mass by subtracting the HAP VOC mass. And NOINTEGRATE means both criteria VOC mass and HAP VOC mass are treated separately. Both criteria VOC mass and HAP VOC mass are kept; the criteria VOC does not include the HAP so we do not need to subtract to prevent double counting

First, prior to any computations, the run schema mechanism table is initialized with the shared schema chemical mechanism that is specified in the run control file. Then all active toxic species

specified by the user, through the optional toxics file, are removed from the run schema mechanism table. And then **all** user specified toxic species, whether active or tracer, are inserted in the run schema mechanism table with a moles_per_mole value of 1.0. For example, if benzene is specified as an active toxic specie, it is removed from the non-toxic portion of the chemical mechanism assignments and completely allocated to the explicit toxic specie. If, on the other hand, benzene is specified as a tracer toxic, it remains in the chemical mechanism assignment (for CBIV: this includes PAR and NR) and also allocated to the explicit toxic specie (BENZ).

The run schema profile weights table is initialized with either the shared schema profile weights or with the optional user specified profile weights. Then, all user specified HAPS species are removed from the run schema profile weights table. If the run type is INTEGRATE then the profiles weights are renormalized following the removal of HAPS, otherwise, they are not renormalized. The profile no longer sums to 1.

A brief outline of the methodology follows.

1. a) Calculate moles model species per gram of emissions

   tmp_calcs_byspc.moles_per_gram
   = tmp_prfwts.percent/100 * tmp_mechanism.moles_per_mole / tbl_species.volatile_mw

   b) Sum over specieId for each mechanism, profileId, aqm_poll

   tmp_calcs_byaqm.moles_per_gram = SUM[tmp_calcs_byspc.moles_per_gram]

2. Calculate mole fractions

   tmp_prfwts.moles = (tmp_prfwts.percent / 100.) / tbl_species.volatile_mw

   tmp_prfwts.moles_pct = tmp_prfwts.moles / SUM(tmp_prfwts.moles for profileId)

3. Calculate average molecular weight of each species in profile
   a) moles model species per moles emissions

   tmp_calcs_byspc.moles_per_mole_em
   = tmp_prfwts.moles_pct * tmp_mechanism.moles_per_mole

   b) Sum over specieId for each mechanism, profileId, aqm_poll

   tmp_calcs_byaqm.moles_per_mole_em = SUM[tmp_calcs_byspc.moles_per_mole_em]

   c) calculate weighting factor

   tmp_calcs_byspc.mole_wtpct
   = tmp_calcs_byspc.moles_per_mole_em / tmp_calcs_byaqm.moles_per_mole_em

   d) calculate average molecular weight on a species basis

   tmp_calcs_byspc.avg_mw = tbl_species.volatile_mw
   * tmp_aqm_carbons.num_carbons / tmp_species_carbons.num_carbons
   * tmp_calcs_byspc.mole_wtpct

   e) calculate average molecular weight by for AQM lumped compound

   tmp_calcs_byaqm.avg_mw = SUM(tmp_calcs_byspc.avg_mw)

4. Calculate values for splits GSPRO file

   tmp_gspro.split_factor = tmp_calcs_byaqm.moles_per_gram * tmp_calcs_byaqm.avg_mw
   tmp_gspro.divisor = tmp_calcs_byaqm.avg_mw
   tmp_gspro.mass_fraction = tmp_calcs_byaqm.moles_per_gram * tmp_calcs_byaqm.avg_mw

5. Calculate data for VOC-to-TOG GSCNV file
   The field **non_voctog** in ***shared.tbl_species*** (imported from SPECIATE 4.0) determines if a compound is considered to be a non-volatile organic gas.

   sumTOG = SUM(tbl_profile_weights.percent)
   sumVOC = SUM(tbl_profile_weights.percent) for tbl.species.non_voctog = FALSE
   tmp_gxcnv.cnv_factor = sumTOG / sumVOC

Output Files from the Speciation Modeling Data Tool

The speciation profile file output from the Speciation Modeling Data Tool contains the split factors, divisors and mass fractions required for speciating the emission inventory criteria pollutants to Air Quality Modeling lumped compounds. The speciation profile file is ready to be used in SMOKE as a GSPRO file. In addition, the VOC-to-TOG conversion factor file is generated to be used in SMOKE as a GSCNV. Note – this file format has been modified from earlier versions. The conversion factors, like the split factors, are specified by profile code, where in previous versions these factors were specified by SCC. Tables 14 and 15 are samples of the Speciation Modeling Data Tool output files.

Table 1. Example of a Surrogate Tool Global Control Variables File

VARIABLE	VALUE	DESCRIPTION
GENERATION CONTROL FILE	/srgtool/surrogate_generation.csv	File containing surrogates for computation
SURROGATE SPECIFICATION FILE	/srgtool/surrogate_specification.csv	File containing settings for generating surrogates
SHAPEFILE CATALOG	/srgtool/shapefile_catalog.csv	Shapefile names and map projection information
SHAPEFILE DIRECTORY	/emiss_shp2003/us	Directory containing all shapefiles needed
SURROGATE CODE FILE	/srgtool/surrogate_IDs.txt	List of surrogate codes and names
SRGCREATE EXECUTABLE	/srgtool/srgcreate.exe	Location of srgcreate executable
SRGMERGE EXECUTABLE	/srgtool/srgmerge.exe	Location of srgmerge executable
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	US36KM_148X112	This is a grid name for regular grid output area
GRIDDESC	/srgtool/GRIDDESC.txt	It is the file containing the list of available of grids (needed only for SMOKE surrogates)
OUTPUT_FILE_ELLIPSOID	SPHERE	Output grid projection ellipsoid
OUTPUT DIRECTORY	/outputsrg/US36KM_148X112	Directory for individual surrogate files
OUTPUT SURROGATE FILE	/outputsrg/US36KM_148X112/srg_total.txt	Name and path for the final merged surrogate file output from srgtool
OUTPUT SRGDESC FILE	/outputsrg/US36KM_148X112/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	srgRun_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. Example of a Surrogate Tool Shapefile Catalog File.

SHAPEFILE NAME	DIRECTORY	ELLIPSOID	PROJECTION	SHAPE TYPE	DESCRIPTION	DATA SOURCE
county_pophu02	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	proj=lcc,+lat_1=33,+lat_2=45,+lat_0=40,+lon_0=-97	Line	Class 1-3 and unknown classified railroads	Transportation Atlas Data & Census 2000 TIGER data
us_lowres	/emiss_shp2003/us	SPHERE	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	/emiss_shp2003/us	SPHERE	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 3a. Example of the Left Columns of the Surrogate Specification File

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 3b. Example of the Right Columns of the Surrogate Specification File

REGION	SURROGATE	Cols 3-9	MERGE FUNCTION	SECONDARY SURROGATE	TERTIARY SURROGATE	QUARTER NARY SURROGATE	DETAILS
USA	Population	…					Total 2000 population
USA	Urban Population	…		Population			Total urban population
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
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 – in bgpro 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.
USA	Total Agriculture	…		Rural Land Area	Land		Sum of: Pasture/Hay, Grains, Row Crops, Fallow Land, Orchards/Vineyards
USA	Industrial Land	…		Urban Population	Land	Population	Sum of building square footage: IND1 … IND6
USA	Heavy and High Tech Industrial (IND1 + IND5)	…		Industrial Land	Urban Population	Population	Sum of building square footage from FEMA categories: IND1 + IND5
USA	Forest External	…	../output/ US36KM_20X20/forest.txt	../output/ US36KM_20X20/ pop_100.txt|Population			
NA	Population	…	Population[USA]; Population[Canada]; Population[Mexico]				

Table 4. Example of a Surrogate Generation Control File

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

Table 5. Example of a Grid Description File

```
!  coords --line:  name; type, P-alpha, P-beta, P-gamma, xcent, ycent
'LAT_LON'
1, 0.0D0, 0.0D0, 0.0D0, 0.0D0, 0.0D0
'UTM_10'
5, 10.0D0, 0.0D0, 0.0D0, 0.0D0, 0.0D0
'LAM_40N90W'
2, 30.0D0, 60.D0,-90.D0,-90.D0, 40.D0
'LAM_40N105W'
2, 30.0D0, 60.D0,-105.D0,-105.D0, 40.D0
' '  !  end coords.  grids:  name; xorig,yorig,xcell,ycell,ncols,nrows,nthik
'EPAW36_56X78'
'LAT_LON' , -127.0D0, 26.0D0, 0.5000D0, 0.33333D0, 56, 78, 1
'NEW_YORK'
'UTM_18', 480.0D3, 4440.0D3, 5.0D3, 5.0D3, 58, 46, 1'
'SMRAQ36_72X74'
'LAM_40N90W', -972.D3 , -1728.D3, 36.D3, 36.D3, 72, 74, 1
'DENVER8_34X45'
```

Table 6. A Sample Output of a RegularGrid 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 7. 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 8. A Sample Log File Created by the Surrogate Tool for RegularGrid Output

```
    Running Date: Fri Dec 23 10:19:25 EST 2005
            Main Control CSV File
    GENERATION CONTROL FILE /srgtool/surrogate_generation_grid.csv
    SURROGATE SPECIFICATION FILE    /srgtool/surrogate_specification.csv
    SHAPEFILE CATALOG       /srgtool/shapefile_catalog.csv
    SHAPEFILE DIRECTORY     /emiss_shp2003/us
    SURROGATE CODE FILE     /srgtool/surrogate_IDs.csv
    SRGCREATE EXECUTABLE    /srgcreate.exe
    SRGMERGE EXECUTABLE     /srgmerge.exe
    DEBUG_OUTPUT    Y
    OUTPUT_FORMAT   SMOKE
```

```
      OUTPUT_FILE_TYPE       RegularGrid
      OUTPUT_GRID_NAME       US36KM_148X112
      GRIDDESC       /srgtool/GRIDDESC.txt
      OUTPUT_FILE_ELLIPSOID   SPHERE
      OUTPUT DIRECTORY       /output/US36KM_148X112
      OUTPUT SURROGATE FILE   /output/US36KM_148X112/srg_total.txt
      OUTPUT SRGDESC FILE     /output/US36KM_148X112/SRGDESC.txt
      OVERWRITE OUTPUT FILES  YES
      LOG FILE NAME   srgRun_grid.log
      DENOMINATOR_THRESHOLD   0.0005
      COMPUTE SURROGATES FROM SHAPEFILES  YES
      MERGE SURROGATES  YES
      GAPFILL SURROGATES  YES

                  Get Grid Header For Surrogate Files
      SRGCREATE_OUTPUT>#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
      SUCCESS IN RUNNING THE EXECUTABLE: SRGCREATE

      Run srgcreate.exe to generate surrogate ratios for USA_100: Population
      SRGCREATE_ERROR>WARNING: QA sum for attr 100, county 51153 was 1.08 not 1; last c,r=123,60
      SRGCREATE_ERROR>WARNING: QA sum for attr 100, county 08059 was 0.93  not 1; last c,r=57,58
      …
      Run srgmerge.exe to gapfill surrogate ratios for USA_570: Heavy and High Tech Industrial (IND1 + IND5)
      SRGMERGE_OUTPUT>Running srgmerge Version 1.1, 07/06/2005
      SRGMERGE_OUTPUT>Using input file '/output/US36KM_148X112/temp_files/gapfill_USA_570.txt'
      SRGMERGE_OUTPUT>Processing Input File line 'OUTSRG=Heavy and High Tech Industrial (IND1 +
IND5); GAPFILL=/output/US36KM_148X112/USA_570_NOFILL.txt|Heavy and High Tech Industrial (IND1 +
IND5);/output/US36KM_148X112/USA_505_NOFILL.txt|Industrial Land'
      SRGMERGE_OUTPUT>Processing output for 'Heavy and High Tech Industrial (IND1 + IND5)'
      SRGMERGE_OUTPUT>Srgmerge completed successfully
      SUCCESS IN RUNNING THE EXECUTABLE: SRGMERGE
      Run srgmerge.exe to gapfill surrogate ratios for USA_328: Forest External
      Error: File --- /output/US36KM_148X112/USA_328_NOFILL.txt -- Does Not Exist

                  Surrogate Generation Summary

      USA_100 Population      SRGCREATE Success
      USA_110 Housing SRGCREATE Success
      USA_120 Urban Population       SRGCREATE Success       GAPFILLING Success
      USA_150 Residential Heating - Natural Gas       SRGCREATE Success       GAPFILLING Success
      USA_240 Total Road Miles       SRGCREATE Success       GAPFILLING Success
      USA_200 Urban Primary Road Miles       SRGCREATE Success       GAPFILLING Success
      USA_255 0.75 Total Roadway Miles plus 0.25 Population   MERGING Success GAPFILLING Success
      USA_340 Land    SRGCREATE Failed
      USA_400 Rural Land Area SRGCREATE Success       Creating GAPFILL TXT File Failed
      USA_310 Total Agriculture       SRGCREATE Success       GAPFILLING Success
      USA_505 Industrial Land SRGCREATE Success
      USA_570 Heavy and High Tech Industrial (IND1 + IND5)    SRGCREATE Success       GAPFILLING
Success
      USA_328 Forest External Creating MERGE TXT File Failed  Creating GAPFILL TXT File Failed
      NA_100  Population      Creating MERGE TXT File Failed
      FINISH -- The Program Run Finished with Some Errors
```

Table 9. The init_db script used to initialize the Speciation Modeling Data Tool

```
#
#  This initialization file needs to be once per installation                    ---
#  Rerunning will delete the existing installation and install the latest version ---
#  of the Speciation Modeling Tool and the shared raw data.                       ---

#  Make a global change to the database name "sptoolv1_2" if you want a different
#  name for the database.

date

dropdb sptoolv1_2
createdb sptoolv1_2
createlang plpgsql sptoolv1_2
psql -c "create schema shared" sptoolv1_2
psql -c "grant create on database sptoolv1_2 to public" sptoolv1_2
psql -c "grant all on schema shared to public" sptoolv1_2
psql -q -f src_model/drop_table.sql sptoolv1_2
psql -q -t -f src_model/table_defs.sql sptoolv1_2
psql -q -t -f src_model/table_inps.sql sptoolv1_2
psql -q -t -f src_model/make_splits.sql sptoolv1_2
psql -q -t -f src_model/prep_out.sql sptoolv1_2
perl src_model/import_rawdata.pl sptoolv1_2 carbons
                  import_data/carbons_cbiv_saprc99.csv
perl src_model/import_rawdata.pl sptoolv1_2 gas_profiles
                 import_data/gas_profile_forimport.txt
perl src_model/import_rawdata.pl sptoolv1_2 species
                 import_data/species_export_wcarter.csv
perl src_model/import_rawdata.pl sptoolv1_2 pm_profiles
                 import_data/pm_profile_export_ascii.txt
perl src_model/import_rawdata.pl sptoolv1_2 pm_profile_weights
                 import_data/pm_species_export_ascii.txt
perl src_model/import_rawdata.pl sptoolv1_2 gas_profile_weights
                 import_data/gas_species_export_nodups_ascii.txt
perl src_model/import_rawdata.pl sptoolv1_2 mechanism
                 import_data/saprc_mechanism_forimport.txt
perl src_model/import_rawdata.pl sptoolv1_2 mechanism
                 import_data/cbiv_mechanism_forimport.txt
perl src_model/import_rawdata.pl sptoolv1_2 static import_data/static.csv
perl src_model/import_rawdata.pl sptoolv1_2 process import_data/process_poll.csv
perl src_model/import_rawdata.pl sptoolv1_2 pm_species import_data/pm_species_xref.csv
```

Table 10. A sample run control file for the Speciation Modeling Data Tool

```
#This is an example test case for the Speciation Tool
#The run control file can have any lines of comment which
#will be ignored on input.  The only lines that are
#imported must have the keyword, followed by a comma ",",
#and the appropriate value specified.

mech_basis,CBIV
run_type,integrate
pro_file,user_prof_wts_2540.txt
output,VOC
tox_file,toxics_active_tracer.csv
haps_file,haps_benz_ethb.csv
splits_out,../split_cnv/prof_intat.out
cnv_out,../split_cnv/cnv_intat.out
```

Table 11. A sample toxics file for the Speciation Modeling Data Tool

```
CBIV,302,BENZ, 6.,A
CBIV,449,ETHBENZ, 8.,T
```

Table 12. A sample HAPS file for the Speciation Modeling Data Tool

```
CBIV,302,BENZENE
CBIV,449,ETHBENZ
```

Table 13. A sample weights profile file for the Speciation Modeling Data Tool

```
TST2540,282,5.72
TST2540,302,2.96
TST2540,438,1.77
TST2540,449,1.23
TST2540,452,7.03
TST2540,508,6.04
TST2540,592,3.14
TST2540,605,2.89
TST2540,620,1.32
TST2540,671,0.84
TST2540,678,2.51
TST2540,717,6.08
TST2540,2200,4.14
TST2540,2284,54.33
```

Table 14. A sample speciation profiles file (GSPRO) from Speciation Modeling Data Tool

```
0000            BENZENE         BENZ        1.00000027  78.11184000   1.00000027
0000            ETHBENZ         ETHBENZ     0.99999998 106.16500000   0.99999998
0000            ETHBENZ         PAR         0.12500000  13.27062500   0.12500000
0000            ETHBENZ         TOL         0.87499999  92.89437500   0.87499999
TST2540         NONHAPTOG       ALD2        0.01704131  32.99254327   0.01704131
TST2540         NONHAPTOG       ETH         0.07337444  28.05316000   0.07337444
TST2540         NONHAPTOG       FORM        0.00004998  16.49627164   0.00004998
TST2540         NONHAPTOG       OLE         0.03663635  30.43771380   0.03663635
TST2540         NONHAPTOG       PAR         0.50270839  15.62038150   0.50270839
TST2540         NONHAPTOG       TOL         0.09527041  98.80567974   0.09527041
TST2540         NONHAPTOG       UNR         0.17751876  15.62097230   0.17751876
TST2540         NONHAPTOG       XYL         0.09739918 115.53876835   0.09739918
```

Table 15. A sample VOC-to-TOG conversion factors file (GSCNV) from Speciation Modeling Data Tool

```
VOC             TOG             TST2540         1.01882178
```

CONCLUSIONS

The Surrogate Tool and Speciation Tool are important new additions that substantially ease the creation of spatial and speciation profiles which are critical inputs to SMOKE modeling. Thus, the tools will reduce the burden of applying SMOKE to new modeling studies. In the future, these tools could be enhanced to create input files in a form that could be used for other emissions models, if desired. The input and output files for both tools are ASCII files that can be managed and versioned by the Emissions Modeling Framework.

BIBLIOGRAPHY

Eyth, A., and W. Benjey. "Generating Sophisticated Spatial Surrogates Using the MIMS Spatial Allocator". In *Proceedings, Transforming Emission Inventories – Meeting Future Challenges Today*; U.S. Environmental Protection Agency: Las Vegas, NV, 2005.

Eyth, A. "The MIMS Spatial Allocator: A Tool for Generating Emission Surrogates without a GIS". In *Proceedings, Emission Inventories: Applying New Technologies*; U.S. Environmental Protection Agency: San Diego, CA, 2003.

Eyth, A., and L. Ran. The Surrogate Tool User's Guide. University of North Carolina at Chapel Hill, 2006. http://www.cep.unc.edu/empd/projects/mims/spatial/srgtool.

Jimenez, M., G. Yarwood, S. Rao. The Speciation Modeling Data Tool User's Guide. ENVIRON International Corporation, 2006.

Vukovich, J. "Implementation of BEIS3 within the SMOKE Modeling Framework". In *Proceedings, Emission Inventories: Partnering for the Future*; U.S. Environmental Protection Agency: Atlanta, GA, 2002.

ACKNOWLEDGMENTS

The work described herein was funded under US EPA contract number 68D-02-066. Software development support was provided by Ms. Limei Ran and Mr. Rajasooriyar Partheepan of the Carolina Environmental Program, and Ms. Michele Jimenez and Ms. Sunja Rao of ENVIRON International Corporation.

KEYWORDS

BEIS3
BELD3
Geographic Information System (GIS) Alternative
Java
PostgreSQL
SMOKE
Spatial Surrogates
Speciation
SPECIATE