Project Owner	Rob de Ligt
Science Maintainers	Rob de Ligt Juha Metsaranta Cris Brack Marcela Olguin
Science Editors	Rob de Ligt
Science Review Panel	Stephen Roxburgh
Science Approval	Werner Kurz
Code maintainers	Max Fellows Malcolm Francis James Leitch
Code Coach	
Documentation Lead	
Release Approval	Werner Kurz

PLEASE READ THIS BEFORE YOU CONTINUE:
This is an open source document. Everybody is welcome to contribute. To allow us to be efficient we have only a few rules:

- All contributors use comments to ask questions or make substantive and structural suggestions. Always suggest solutions in your comments not problems.
- Only Maintainers can resolve comments.
- All contributors make edits in "suggesting mode" (even if you are a maintainer or editor, so the other maintainers or editors can cross-check and accept your contribution.)
- Write suggested edits directly into the text in 'suggesting mode' do NOT use comments to suggest edits
- Only Editors assigned to the document can approve edits
- Spelling will be UK English
- Referencing Style is Harvard (author-date) type. Tools such as Zotero may be useful.

More info about how the moja global community works on projects can be found here.

Design of sampling and aggregation method for uncertainty estimation in the Full Lands Integration Tool (FLINT)

February 15, 2020

Abstract

Uncertainty assessment is important for land based greenhouse gas estimation, as it is incorporated in all aspects of reporting, from national greenhouse gas inventories through to results-based payments for REDD+. The purpose of this design document is to describe the changes to the FLINT system for selection of land units and aggregation. This design will provide:

1. Efficient methods for first selecting the land units on which Monte Carlo analysis will be carried out and;
2. Methods for aggregating the results of Monte Carlo simulations to allow for estimation of uncertainty for a simulation. The method will not be fixed to any single implementation allowing for scalability and use by multiple users.

This work is linked to the Monte Carlo process provided by the FLINT. This document focuses on the design of the sampling method of the land-units that need to be processed through a Monte Carlo analysis and the aggregation of the Monte Carlo results into uncertainty estimates.

A key design requirement is that the methods are consistent with the 2006 IPCC Guidelines for national greenhouse gas inventories, and are also capable of supporting other common land sector MRV programs and reporting requirements.

Introduction

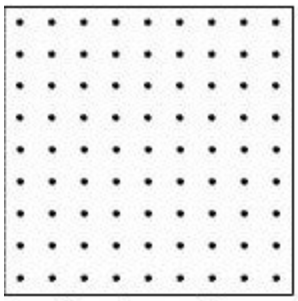
With the addition of Monte Carlo analysis capability in the FLINT framework there is now the possibility to run N possible outcomes of a simulation based on uncertainty variables defined by know distributions. There is the requirement to derive a smaller representative sample of large datasets, due to the computational cost and extremely large result sets generated from these repeated simulations. The FLINT will run the Monte Carlo method on these smaller samples and will require the ability to aggregate the results from this new type of simulation output.

Sampling

The FLINT will be modified to produce a new raster layer from a study region dataset. Once the spatial inputs are defined by a provider file the FLINT will analyse these layers and output a sample inclusion raster. This raster can then feed into a Monte Carlo simulation run. The raster will be a simple mask that defines if a cell is included in the monte carlo run. The user will be able to select the type of sampling and the number of samples required. The FLINT will be extended to provide the following sampling methods: Systematic Uniform Sampling, Simple Random Sampling, and Stratified Random Sampling . The monte carlo simulation will also be able to include any boolean raster to provide the cells simulated in a monte carlo run. These rasters can also be produced by existing GIS software tools. The raster should just use the values 1 and 0. 1 to indicate a sample included in the run and a 0 to exclude the cell.

Systematic Uniform Sampling

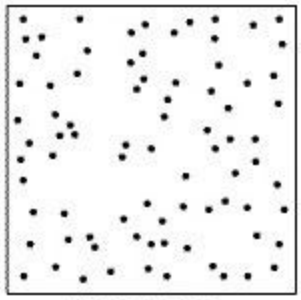
The simplest approach will just take samples at a regular intervals from the simulation area. This could be used for a simple regular stratified simulation area but will not provide much benifit for more complex landscapes. This approach can be very simply built as part of the other approaches.



Systematic Sampling

Simple Random Sampling

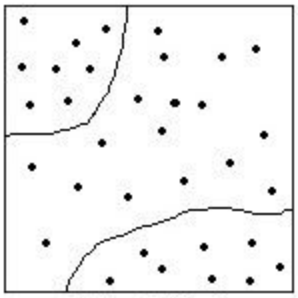
The next simplest approach is randomly choosing sampling points within the simulation area, the points are simply randomly distributed throughout the area.



Random Sampling

Stratified Random Sampling

The last approach is probably the most useful in a FLINT context. The flint will take the input spatial data for a simulation and produce thematic classes from the data. A number of random points will be allocated based on these classes Then the FLINT will stratify the number of random points so that larger classes will receive more points, this ensures classes that cover a large area will receive more sampling points, but it also ensures smaller area classes will still have sampling points.



Stratified Sampling

FLINT extensions for sampling

The FLINT will be extended to provide the ability to create Spatially Balanced Cell Samples. The extension will implement the Reversed Randomized Quadrant-Recursive Raster (RRQRR) algorithm (Theobald et al. 2007). This algorithm is raster based and a good fit for the FLINT system. The algorithm requires an inclusion probability layer that provides the probability of a cells inclusion in the simulation relative to others. This inclusion layer can be generated outside the system and the ability to generate it from a configuration with also be provided. The RRQRR algorithm is used to map 2D space into a linear distribution of samples in which successive samples constitute a spatially balanced simulation design. (Theobald et al. 2007) recommends the number of samples drawn from the distribution be less that 1 percent of the simulation area.

Probability Inclusion generation

The FLINT will produce an inclusion probability layer for a simulation configuration. This layer will then be used as an input to create sample cells for a Monte Carlo run. The sample cells layer is a simple boolean layer indicating inclusion(1) or exclusion(0) of a cell in the monte carlo simulation.

The inclusion generation is driven from an inclusion strata config and a provider config. The inclusion strata config defines layers from the provider file that defines the strata for the simulation. The strata will then be used to produce the inclusion probability raster based on the defined layers. An example probability inclusion config is shown below:

```
{
  "inclusion_strata": [
    {
      "layer": "plantations phe",
       "type": "discrete"
     },
     {
       "layer": "forest_cover",
       "type": "time_series"
      },
      {
        "layer": "forest_type",
         "type": "discrete"
      },
      {
        "name": "primary_forest_extent",
        "type": "time_series"
       }
   ]
}
```

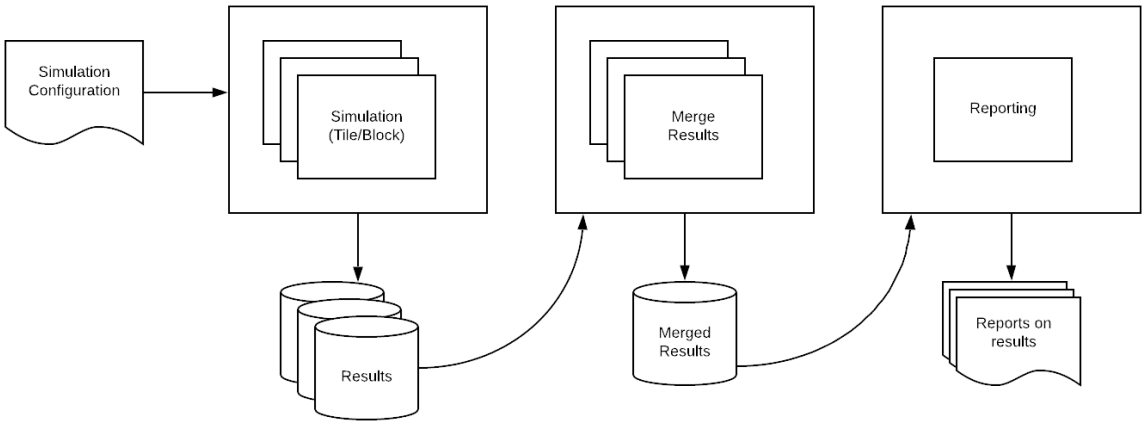
The FLINT is then run in inclusion layer creation mode to create a raster of cell inclusion probabilities.

Sample cells generation

The sample cells generation can be run in one of three ways: Systematic Uniform, Simple Random or Stratified Random. This process takes as input the simulation configuration and the inclusion probability raster. Out of this a boolean mask raster layer is created that indicates the cells that will passipitate in an Monte Carlo simulation run.

Aggregation

The FLINT framework has the ability to process enormous amounts of information for a single simulation. To achieve usable runtimes of these simulation the FLINT breaks the problem down into smaller chunks and distributes the work amongst multiple computers. With this distribution comes the requirement to aggregate the results back into a single simulation result set



Distribution of processing

As the FLINT process cells through simulation models it records the movements(fluxes) between pools. An example flux record in JSON format is shown below.

```
{
   "monte_carlo_flux_record": {
       "id": 10101,
       "year": 1999,
       "source_pool": "atmosphereCM",
       "destination_pool": "soilOrganicCM",
       "flux_value": 1.194643,
       "lu_count": 9846,
       "iteration": 756
   }
}
```

This information is recorded for every flux generated by any model running in the system at the completion of the cell simulation the results are aggregated into a the collection of fluxes for the simulation unit being processed (typically a tile block) when the block is complete these aggregated fluxes are then written to storage. When all simulation units are processed they themselves are aggregated into a single simulation result. With the addition of Monte Carlo simulation capability the results from a single cell has changed from single output values to stochastic sets of N values that represent the result for N possible simulations given defined distributions of configured inputs. When it comes time to produce reports from these aggregated flux sets they can be applied to produce a table of outputs for a given time period and stepping. The emerging discipline of probability management communicates uncertainties as arrays of trials called SIPs, which may be added up across simulations like numbers [4]. SIPs of various outputs (generated on multiple platforms at multiple levels) may be added, multiplied or used in other calculations, and then rolled up to higher levels.

Calculation of Uncertainty Statistics

When the results of a simulation are recorded as arrays of possible results, statistics can be derived from these arrays. The FLINT will report uncertainties as percent uncertainty relative to a particular two-tailed confidence interval. The simulation configuration will allow specification of reporting confidence levels (e.g., 90%, 95%) . The percent uncertainty will be calculated relative to the specified confidence level. It will be possible to configure the reporting of the confidence intervals and not just the percent uncertainty.

Moja command line executable

The command line tool (CLI) will require some extensions to handle the new generation and aggregation functions . The CLI will need to be extended to run these new functions

The current command line options are:

```
Allowed options:

General options:
  -h [ --help ]          produce a help message
  --help-section arg     produce a help message for a named section
  -v [ --version ]       output the version number

Commandline only options:
  --logging_config arg   path to Moja logging config file
  --config_file arg      path to Moja run config file
  --provider_file arg    path to Moja data provider config file

Configuration file options:
  --config arg           path to Moja project config files
  --config_provider arg  path to Moja project config files for data providers
```

The proposed changes will add the following.

```
Allowed options:

General options:
  -h [ --help ]          produce a help message
  --help-section arg     produce a help message for a named section
  -v [ --version ]       output the version number

Commandline only options:
  --logging_config arg   path to Moja logging config file
  --config_file arg      path to Moja run config file
  --provider_file arg    path to Moja data provider config file
  --inclusion_config_file       path to Moja inclusion probability config file


Configuration file options:
  --config arg           path to Moja project config files
  --config_provider arg  path to Moja project config files for data providers
```

References

1. Design for Monte Carlo and Propagation of Error Uncertainty in the Full Lands Integration Tool
2. Stevens, D.L., and A.R. Olsen. 2004. "Spatially balanced sampling of natural resources."Journal of the American Statistical Association 99 (465): 262–278
3. Theobald, D.M., D.L. Stevens, Jr., D. White, N.S. Urquhart, A.R. Olsen, and J.B. Norman. 2007. "Using GIS to Generate Spatially Balanced Random Survey Designs for Natural Resource Applications."Environmental Management 40: 134–146.
4. Sam Savage, Stefan Scholtes and Daniel Zweidler, 2006, “Probability Management,” *OR/MS Today*, February 2006, Vol. 33, No. 1.