nFIESTA (new Forest Inventory ESTimation and Analysis)

Functional description of the software

Ivo Kohn[*], Radim Adolt[*], Jiří Fejfar[*], and Adrian Lanz[†]

[*] *Forest Management Institute Brandýs nad Labem (ÚHÚL). Czech Republic.*
[†] *Federal Institute for Forest, Snow and Landscape Research WSL, Birmensdorf. Switzerland.*

Task T2.3.1 Common data analysis and output delivery system

---

updated on February 6, 2019

Contents

1 Introduction

The new Forest Inventory ESTimation and Analysis system (nFIESTA) is a software solution capable to produce sound estimates of target parameters describing the condition of forests and its development. It has been developed as part of the Diabolo project (http://diabolo-project.eu/) funded by the European Union's Horizon 2020 research and innovation programme.

The specificity of the nFIESTA estimation system is that it allows for an analysis of large datasets sourced by National Forest Inventories (NFIs) of many (not only European) countries. Such an integrated database contains sample plot data that use to be collected following a number of different sampling and survey designs, with spatial distribution and timing that vary from country to country. The only methodological prerequisite is that the NFI (field) data must come from a probability sample.

Using such a system the estimates of forest and landscape parameters can be obtained at a level of detail starting from a relatively small areas - as small as several hundreds of sq. kilometers - for which (in a rather extreme case) even no sample plots need to be available. From the temporal point of view, the current implementation supports estimations for periods of one calendar year or longer. To make all this possible the system was designed and developed so it can use various sources of auxiliary information (typically GIS maps produced by remote sensing methods).

The technological solution of nFIESTA has been evolving over several years. In 2017 the first prototype was made operational. At that time artificial plot data that mimicked the field reality (various GIS maps depicting the presence of forest and its predominant type) were used to boost the development. Next development iteration was connected to a case-study within the task T2.3.1 of the Diabolo project, which already used real NFI data from four countries - Czech Republic, France, Germany and Switzerland. Each of the four countries made available plot data covering the whole country's extent (a probability sample) with attributes related to the above ground tree biomass. As a source of auxiliary data Copernicus Forest Type 2012 and Tree Cover Density 2012 maps were used. This integrated dataset has been extensively analysed by the nFIESTA which provided a great benchmark for further development. During this case-study the software and also the estimation methods were intensively fine-tuned.

The nFIESTA is build on top of the PostgreSQL RDBMS (https://www.postgresql.org/). It is distributed and maintained in the form of a standardised PostgreSQL Extension mechanism (https://www.postgresql.org/docs/11/extend-extensions.html). The only software prerequisites are the availability of the PostgreSQL database with the PostGIS extension (both free and open source software). So far no graphic user interface (GUI) exists for nFIESTA system - except the common front ends used to access the PostgreSQL database e.g. PgAdmin (https://www.pgadmin.org), which is

also a free and open source application. For the time being there are no plans to extend the nFIESTA by a GUI. But of course, this attitude may change depending on the possibilities and needs of further development.

The nFIESTA is distributed under the license EUPL and the source code is hosted and accessible at GitLab repository(https://gitlab.com/nfiesta/nfiesta_pg), which is publicly available. Standard installation instructions can be found in a README.md file (https://gitlab.com/nfiesta/nfiesta_pg/blob/master/README.md). In the following main functional components of nFIESTA are described.

2 Functional overview of nFIESTA

The basic logic of the nFIESTA application can be described by the schema in figure 1.

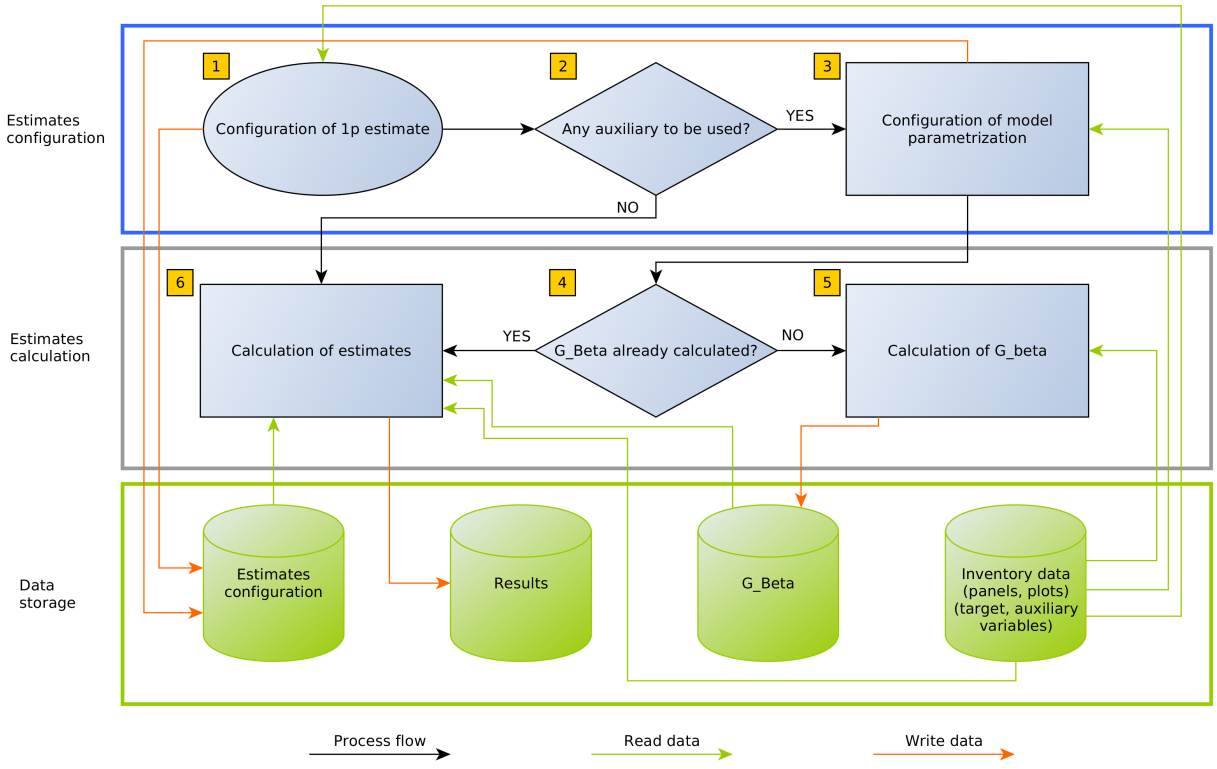

Figure 1: Functional schema of the nFIESTA estimation software.

It is obvious that before any estimate can be produced there must be at least some plot data uploaded to the nFIESTA database. The NFI as well as auxiliary data can be sourced to nFIESTA using generic ETL mechanism.

Estimates of target parameters can be produced in a chain of six interlinked processes shown by the functional schema and described below:

1. **Node No. 1** - the user configures a one-phase estimation task (no auxiliary data is considered for one-phase estimation). The configuration itself consists in the specification of the estimation cell(s), time frame (in terms of one or several calendar years) and the target variable. The user can choose from a list of options reflecting the current content of the nFIESTA database.

2. **Node No. 2** - the user decides whether any (wall-to-wall) auxiliaries can and should be used to get more accurate estimates. This decision can hardly be automated as the stem 'does not know' what auxiliaries could be useful for what target variables. It is a sole judgment of the nFIESTA user.

3. **Node No. 3** - (after deciding to use auxiliaries) the user has to specify which auxiliaries will be used (explanatory variables), how the (linear) working model will be specified and what region(s) will be used for the working model parametrization. The choice of parametrisation region D_+ is generic in the sense that the user only chooses the type (or level) of parametrisation region. All parametrisation regions are defined in terms of unions of estimation cells and the particular type of parametrisation region represents a set of spatially congruent parameterisation regions. Once the type of parametrisation regions is chosen the system can automatically identify a parameterisation region for each of the considered estimation cells. The specification of auxiliaries and the working regression model needs to be done by the user.

4. **Node No. 4** - nFIESTA automatically checks whether the $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ matrices are already available for the combination of parametrisation region, set of auxiliaries and working model. The beta-matrices are used to estimate coefficients of the working model in a parametrisation region in an efficient way. They depend on auxiliaries and working model specification but they are independent on the choice of the target variable. Because potentially many estimation tasks may use the same auxiliaries and working model but different target variables, the nFIESTA strategy of storing once calculated beta-matrices saves potentially many (unnecessary) computations. Consequently the requested estimators can often be produced in a (much) shorter time. Further details on the construction of the $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ can be found in a technical report by Adolt *et al.* [2018, p. 21].

5. **Node No. 5** - nFIESTA calculates the beta-matrices for each parametrisation region (based on the working model specification). The list of

parameterisation regions for which these matrices need to be evaluated is determined based on the linkage of cells and parametrisation regions of the selected type (see node No. 3). Finally the matrices are stored in the database so they can be found and used for similar configurations (but different target variables).

6. **Node No. 6** - if the overall process reaches this node all necessary data and metadata (estimate configuration) is passed to the estimation functionality so the estimates are produced and stored in the corresponding part of the database. This node can be reached either directly from node No. 2 (one-phase estimation using only field plot data), or via node No. 4 (estimators which use auxiliaries, beta-matrices already available) or through node No. 5 (estimators which use auxiliaries, beta-matrices have just been calculated).

As it can be seen on figure 1 nFIESTA can be broken down in to three logical parts: i) Estimates configuration, ii) Estimates calculation and iii) Data storage. These will be described in more detail in the three sections which follow.

2.1 Data storage

Figure 2 shows relational structure of the nFIESTA database. It is a detailed representation of the lower part of the functional schema shown in figure 1.

The left part of the data storage schema, labeled as Inventory data, was designed to contain data provided by National Forest Inventories i.e. field plot data and all necessary design and survey metadata. In addition the Inventory data part also stores the linkage of inventory plots to (predefined) estimation cells as well as auxiliary values determined at (exact) inventory plot positions. The reason behind is that many countries are not allowed or prefer not to disclose exact coordinates of their (usually permanent) National Forest Inventory plots.

The Inventory data part is supposed to be filled by an ETL (Extract Transformation Load) process working as a bridge between the nFIESTA and any system facilitating the process of NFI provision by countries (e.g. the future E-Forest 2.0).

The right part of the schema, called Analytical part, is used by nFIESTA to configure, produce and save estimates of the requested target parameters.

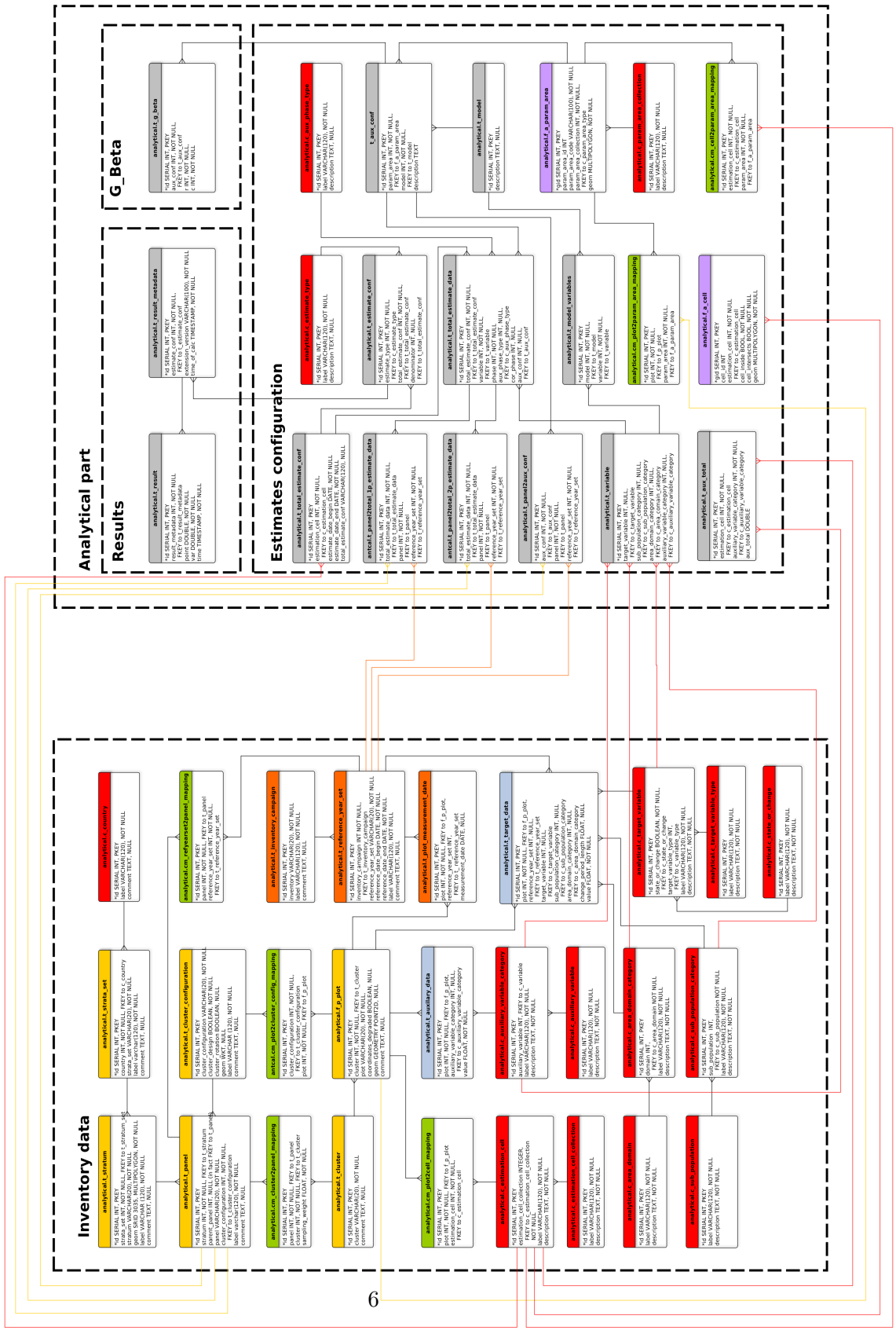

Figure 2: Relational schema of database part of the nFIESTA.

Analytical part can be broken down into three subparts:

1. **Results** - to the two tables of this subpart all the estimates produced by nFIESTA and their metadata are saved. Through the metadata the estimates keep a link to the their configuration so any time later the user can see how the particular estimates were produced. Results are written by the process described in node No. 6 of the functional schema in figure 1.

2. **G_Beta** - this subpart contains only one table in which the precalculated $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ matrices are saved. The beta-matrices can be used to estimate new target variable applying a set of auxiliaries and working model which had been previously used for one or more other target variables. Further information on this matrices can be found under the description of node No. 4 of the schema in figure 1.

3. **Estimates configuration** - here all metadata concerning the estimates configuration are stored. The nFIESTA Estimates calculation functionality accesses this subpart in order to steer the estimation process (prepare the right data, auxiliaries and run the appropriate estimation procedure). The process of estimates configuration is briefly described in nodes No. 1 to 3. of the functional schema in figure 1.

The Data storage schema uses different colors to highlight certain sets of database tables according to their meaning. The meaning of the colors is the following:

Red - code lists (or lookup tables) defining the key categorial variables recognised by the system e.g. list of countries (analytical.c_country) or the list of target variables (analytical.c_target_variable). Data provided to the system refer to categories found in these code lists. This way semantical meaning as well as a proper structure of inventory data being uploaded can be enforced. These code lists are predefined by the nFIESTA user (administrator).

Yellow - metadata describing spatial aspects of the particular NFI sampling and survey design.

Orange - metadata describing temporal aspects of the particular NFI sampling and survey design.

Blue - actual data provided by NFIs at the plot level. The data must be linked (enforcement by foreign keys) to the code lists (red tables) as well as design and survey metadata provided by the respective NFI (country).

Green - mapping tables which are used to model more complex (N:N) relationships between tables. These are mostly found between tables describing the particular NFI design but a good example is the analytical.cm_plot2cell_mapping, which makes it possible to link one plot to several (spatially overlapping) estimation cells.

Violet - these tables contain spatial representation (geometries) of estimation cells (study areas for which estimates need to be produced (analytical.f_a_cell) and of parametrisation regions (.f_a_param_area) formed as unions of available estimation cells (in order to preserve the spatial linkage to plots via the linkage between plots and cells established by NFIs).

Gray - used for tables in the Analytical part of the nFIESTA Data storage. These tables are filled during the production of estimates based on the content of the Inventory data part.

The meaning and intended use of particular data tables of the Inventory data part (left part of the Data storage schema in figure 2) can be found in the technical report by Lanz *et al.* [2018b]. The description of tables used within the Analytical part can be found in table 1 below.

Table 1: List of tables in the Analytical part of nFIESTA Data storage.

Table name	**Description**
c_aux_phase_type	Code table of sampling phases. Reflects various auxiliary data types (wall-to-wall maps, deified sampling grids).
c_estimate_type	Code table of estimate types (total or ratio).
c_param_area_type	Code table of parametrization area types i.e groupings of parametrization areas which are spatially congruent.
cm_plot2param_area_mapping	Mapping table providing the linkage of sample plots to parametrization areas.
f_a_param_area	Table containing parameterisation areas themselves (their geometries).
t_aux_conf	Table with possible combinations of parametrization areas and working models (listing the explanatory i.e. auxiliary variables).
t_aux_total	Table storing (known) totals of auxiliaries (wall-to-wall maps) in the estimation cells.
t_estimate_conf	Table listing existing estimates configurations (configuration of total or combination of two totals in case of ratio).
t_g_beta	Table containing the precalculated $\check{\boldsymbol{G}}_{\beta_{t+}}$ matrices (linked to particular auxiliary configuration).
t_model	Table containing the working regression models (defined sets of explanatory variables, just a container with unique id).

Continued on next page

Table 1 – *Continued from previous page*

Table name	**description**
t_model_variables	Table listing explanatory (auxiliary) variables used within a working model.
t_panel2aux_conf	Table containing sampling panel to working model mappings.
t_panel2total_2ndph_est_data	Table containing sampling panel to field survey data mappings (specific for a particular estimates configuration).
t_panel2total_1stph_est_data	Table containing sampling panel to auxiliary data mappings (specific for a particular estimates configuration).
t_result	Table with the estimation results.
t_result_metadata	Table storing the metadata attached to estimation results (date of calculation, nFIESTA extension version etc.).
t_total_estimate_conf	Table storing configurations of the estimates of totals.
t_total_estimate_data	Table where (most) of the data used for particular estimation task is specified including the sampling phase.
t_variable	Table listing available combinations of variables (target as well auxiliaries), sub-populations and area domains
v_ldsity_conf	Database view on table t_variable with human readable labels.

2.2 Estimates configuration

This part of the nFIESTA functionality formalises the search for the best possible data (field as well as auxiliaries if available) and the corresponding estimation technique which can be used given the user specification of:

- **target parameter** - variable e.g. total forest area
- **estimation cell** - representing a geographical area (denoted by D) for which the best possible estimate of the specified target parameter needs to be identified as well as the dataset allowing for its calculation
- **inventory year(s)** - to which the estimate should (ideally) correspond
- **parametrisation region** - a region $D_+ \supseteq D$ in which a working model can potentially be parametrised and used to increase the precision of target parameter estimate

The configuration process is captured by schema in figure 3. In this schema the orange nodes correspond to user defined inputs (left part of the schema). The central rectangle corresponds to purely design-based estimation, while in the right one leads to model-dependent or mixed (i.e. for some sampling strata model-dependent and design-based for other). Section 6 of the report by Ene *et al.* [2018] provides a better insight in terms of estimation methods to be used for each of the possible paths between nodes No. 1 and 10.

The blue and gray diamonds to decisions made by nFIESTA (potentially user-assisted decision) and the thin blue or gray rhomboids represent dataset available for the calculation of the estimate(s). The only green end point corresponds to a state when the estimate of the respective target parameter can be calculated. Reaching the red end point will stop the whole process because no viable configuration (estimation method) exists for the user inputs.

The estimate configuration logic can be described following the nodes of the schema in figure 3:

- **Node No. 1** - the user chooses the target parameter (total or ratio, target variable), the reference year(s), the estimation cell and parametrisation area type (to be able to test whether auxiliary data can be used for estimation). For a given estimation cell and parametrisation region type (defining a set of spatially congruent parametrisation areas) a particular parametrisation region is easily determined based on an explicit linkage to estimation cells (the unions of which define parametrisation regions).

- **Node No. 2** - automatic check whether there the necessary target variable(s) is (are) available on all plots within the automatically identified parametrisation region of given type. The check is done irrespective of the reference year.

- **Node No. 3** - because the requested target variable(s) is not available for the whole or part of parametrisation region no estimates can be produced. From here the user has always the chance to come back to node No. 1 and choose another type of parametrisation region (possibly the one corresponding to the estimation cell itself) and continue with second iteration of the configuration. In fact such a logic could be implemented automatically, but at the moment it is has not been implemented by nFIESTA.

- **Node No. 4** - Data identified in node No. 4 is reduced to a subset corresponding to the desired reference year(s). Then the size of the subset is compared to the minimum sample size requirement. reference year(s).

- **Node No. 5** - plot data (organised to sampling panels) which fits spatially as well as temporarily to the requested target parameter, cell and parametrisation region.

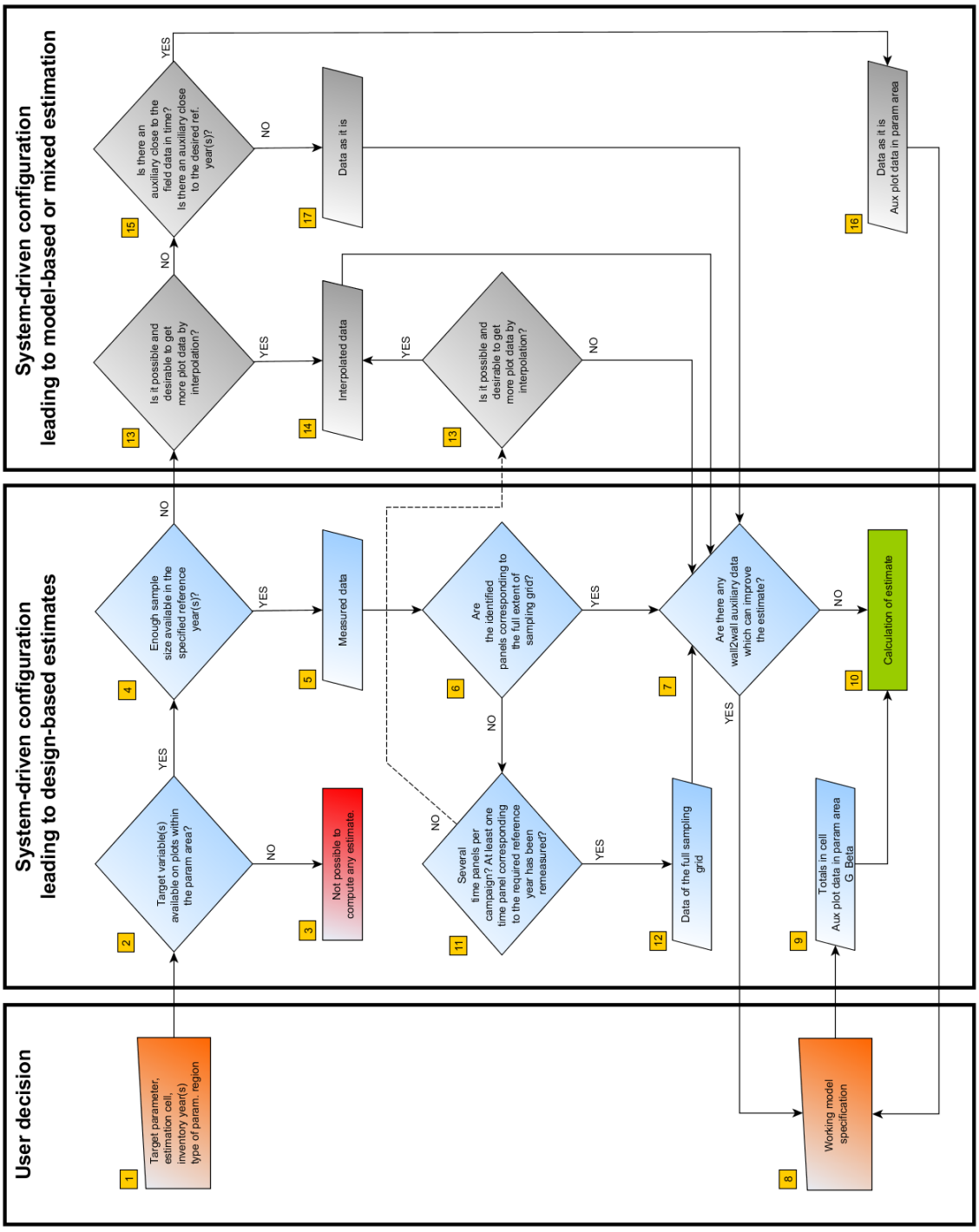

Figure 3: Schema of nFIESTA estimates configuration logic.

- **Node No. 6** - system check whether the found plot data represents the full sampling grid surveyed in the field. The purpose of this check is to separate situations when a single-phase estimation can be performed and no potentially better estimation technique exists (using wall-to-wall auxiliaries is considered also a kind of single-phase estimation here).

- **Node No. 7** - user assisted decision whether auxiliary data obtained from wall-to-wall maps can be used to improve precision of the target parameter estimate(s).

- **Node No. 8** - the user has to specify auxiliaries and all the terms of a linear regression model which will be used for the estimation.

- **Node No. 9** - the dataset necessary for the particular estimator type i.e. $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ matrices and auxiliary totals for the extent of the given estimation cell. The gbeta-matrices can either be reused (if available) or they have to be evaluated by nFIESTA right at this moment.

- **Node No. 10** - nFIESTA calculates the respective estimate the type of which depends on the path through which the node has been reached.

- **Node No. 11** - system checks whether there are more panels representing various reference years in the dataset identified in node No. 5. If yes an additional check is performed to find out whether for each of the reference years at least one panel exists which has been re-measured. If both conditions are met a two-phase estimator can be constructed combining the panels for the requested years and panels which were measured in the past (or future in some cases).

- **Node No. 12** - data corresponding to full sampling grid (all yearly panels not necessarily form the same inventory campaign) which can be used for two-phase estimation (to improve precision, old plot data can be used as non-exhaustive auxiliaries). Wall-to-wall auxiliaries may also enter the game if a positive answer in node No. 7 is obtained. In such a case the resulting estimator would actually be a three-phase one.

- **Node No. 13** - either there was no suitable data found in node No. 5 or the data does not correspond to the full sampling grid and the available panels do not allow for two-phase estimation (node No. 11). In the former case one can try an interpolation of plot values between two measurement years or using the unchanged plot values corresponding to the next preceding survey (if an interpolation between two field survey occasions is not possible). This corresponds to model-based or mixed inference. In the later case (i.e. when reaching node No. 13

from node No. 11) the default system behavior is to continue to node No. 7 (avoiding interpolation at the end a design-based estimation is pursued).

- **Node No. 14** - the system creates a plot dataset which contains the original as well as some artificial panels with interpolated plot data. Currently only a linear interpolation can be used. This can be improved by disturbance maps and growth models (once they are available for large territories). From here the configuration goes to node No. 7 to check whether wall-to-wall auxiliaries can be used to improve precision of the estimate. Estimates using interpolated data should not be considered generally unbiased. The interpretation of the respective variance estimates should reflect this circumstance.

- **Node No. 15** - a situation of no data measured in specified reference year(s) and no interpolation possible. Check if there are is wall-to-wall auxiliary data corresponding to the measurement year(s) of field data identified in node No. 5 and at the same time, if there is another (generation) of the same auxiliary data corresponding to the user specified reference year(s).

- **Node No. 16** - field data identified in node No. 5 and the pair(s) of wall-to-wall auxiliaries corresponding to the field data and the requested reference year(s). Next step is the parametrisation (calculation of G_Beta) based on the auxiliary corresponding to the only available field data followed by a synthetic estimation using the fitted model but auxiliaries corresponding to the user requested reference year(s). This method is not implemented in nFIESTA because of lack of suitable auxiliaries with a regular update frequency and large spatial coverage.

- **Node No. 17** - there is no other option than to take the inventory data identified in node No. 5 as it is and potentially try to improve precision by incorporation of wall-to-wall auxiliaries (timely indifferent approach).

At the moment the assessment of all the above conditions is done in a way that a positive answer is obtained only in case the conditions are met for each sampling stratum encroaching the parametrisation region. This approach basically means that we head toward the same type of estimate in all strata found within the given estimation cell. This constraint can be relaxed in the future implementation but for the moment a more transparent approach was preferred.

2.3 Estimates calculation

This part of nFIESTA implements the estimators described by Adolt *et al.* [2018]. The selection and preparation of data for the particular estimate is done by a set of SQL functions. These are also used to prepare auxiliary data and parametrise the working model whenever a modified direct estimator is specified in the estimates configuration. At certain stage the SQL functions hand over to a PL/Python function which calculates the matrix inversion needed to get $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ matrix.

Once all data for the estimation has been prepared other SQL functions calculate the desired estimates of total or ratio (depending on what has been configured) while another PostgreSQL extension called *htc* (`https://gitlab.com/nfiesta/nfiesta_htc`) is called to evaluate the respective variance estimate according to the Horvitz-Thompson theorem for infinite populations [Cordy, 1993]. This extension also evaluates the pairwise inclusion densities following the specification given in technical report by Adolt *et al.* [2018, p. 6].

For sake of calculation speed the core part of the *htc* extension has been written in C but there is an SQL API (Application Interface) making the integration with PostgreSQL possible. The *htc* extension is automatically installed during the standardised installation of nFIESTA so the user in fact does not have to interact with it directly.

Thanks to the database (SQL) nature of the nFIESTA implementation many target parameters can be evaluated in one single run - for a large number of estimation cells and, if requested, also for many alternative parametrisation regions and working models. The calculations can be distributed to more processor cores or even physical machines so the results are obtained in a fraction of the time, which would be normally needed. Unfortunately, the spread of the calculation load to more cores or machines has to be steered manually at the moment.

The nFIESTA performance was tested within the T2.3.1 case study [Lanz *et al.* , 2018a]. With a PostgreSQL (version 11) and nFIESTA installation on a vmware virtual machine (64bit Windows 7 OS, 2x Intel Xeon 2.00GHz / i.e. 4 cores in total, 8GB RAM, 127 GB SSD disk) 36990 estimates generated by combinations of:

- seven target parameters (total biomass, total coniferous biomass, total broadleaved biomass, total area of biomass domain, mean total, mean coniferous and mean broadleaved biomass per hectare of the biomass domain)
- the 50 by 50 km and 100 by 100 km estimations cells (Inspire grid) covering (or intersecting) the whole territories of Czech Republic, France, Germany and Switzerland, plus one separate cell corresponding to the

whole territory of Czech Republic (to test *htc* extension performance for larger estimation cells)

- 100 by 100 km parametrisation regions plus one separate parametrisation region corresponding to the whole territory of the Czech Republic
- six alternative working models (defined in terms of variables generated from the Copernicus Forest Type and Tree Cover density maps)

were calculated using three parallel PostgreSQL connections (three cores of one physical machine) in less than six hours. Out of this time three hours were spent with the working model parametrisation (calculation of $\tilde{\boldsymbol{G}}_{\beta_{t+}}$ matrices), one hour took the computation of total estimates and one hour and twenty minutes took the computation of ratios (including the calculation of the respective variances by *htc* extension).

References

Adolt, R., Fejfar, J., Lanz, A., & Ene, L., T. 2018. *nFIESTA (new Forest Inventory ESTimation and Analysis) Estimation methods.* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (the EU's Horizon 2020 programme).

Cordy, C. B. 1993. An extension of the horwitz-thompson theorem to point sampling from a continuous universe. *Statistics and probability letters*, **18**, 353–362.

Ene, L., T., Adrian, L., Adolt, R., & Fejfar, J. 2018. *Integrating NFI field data and auxiliaries from various time points to produce an up-to-date information on status and changes of forest resources. A selection of methods proposed for the forest information and estimation system (nFIESTA).* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (DIABOLO).

Lanz, A., Adolt, R., Fejfar, J., Kohn, I., Morneau, F., Pesty, B., & Riedel, T. 2018a. *nFIESTA (new Forest Inventory ESTimation and Analysis) Demonstration study based on NFI plot data and large-area, high-resolution auxiliary data.* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (the EU's Horizon 2020 programme).

Lanz, A., Adolt, R., Kohn, I., & Fejfar, J. 2018b. *Specification of the eFOREST 2.0 CSV file formats.* Tech. rept. Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks (the EU's Horizon 2020 programme).