Deliverable

D2.1 Initial characteristics for mapping and estimation platform

PathFinder Project

Version: 1.0

27 February 2024

The research leading to these results has received funding from the European Union Horizon Europe (HORIZON) Research & Innovation programme under the Grant Agreement no. 101056907

I. DOCUMENT CONTROL

Project	PathFinder (101056907)
Project Title	Towards an integrated consistent European LULUCF monitoring and policy pathway assessment framework
Date	27 February 2024
Author/s	Jukka Miettinen*, Radim Adolt[+], Johannes Breidenbach[‡], Jiří Fejfar[+], Jana Hanáková[+], Ivo Kohn[+], Lukas Kratěna[+], Mari Myllymäki[§], Lauri Seitsonen*, Renne Tergujeff*, Jiří Závodský[+]. *VTT, Contribution to Forestry TEP sections [+]ÚHÚL, Contribution to nFIESTA sections [‡]NIBIO, General overview, advice and supervision [§]LUKE, General overview, advice and supervision
Reviewer/s	Annemarie Bastrup-Birk (EEA), Ambros Berger (BFW), Thomas Riedel (TI).
Activity	WP2, Task 2.1 Developing a mapping and estimation platform
Dissemination Level	PU
Filename	PathFinder_D21-Initial_characteristics_for_mapping_and_estimation_platform.docx

DISSEMINATION LEVEL

PU Public, fully open access

II. DOCUMENT HISTORY

Version	Date	Author	Change
0.1	16.1.2024	See above	Full draft for internal review
1.0	27.2.2024	See above	Reviewer comments integrated

III. IMPORTANT NOTICE

This report often refers to National Forest Inventory (NFI) data, Sentinel data, Copernicus products and 1 km INSPIRE grid cells. Although, these are considered relevant datasets in the European context and are used in the platform demonstration during the project, they are not an integral part of the platform design, and they do not define the platform´s capabilities. Instead of NFI data any other field data fulfilling the definition of probability sample in a geographical space can be processed by the platform. Sentinel data can be changed for any comparable Earth observation data such as Landsat and similar maps can still be produced. Copernicus products can be changed for any other maps that can serve as auxiliaries for estimation with similar benefits. And finally, the 1 km INSPIRE cells can be replaced by another INSPIRE level (10 km, 100 km…), by certain level of NUTS regions or by any other partitioning of geographical space to increase secrecy of coordinates of permanent sample plots. For better readability, we do not emphasize this aspect each time we refer to these datasets in the document.

IV. EXECUTIVE SUMMARY

The PathFinder project aims to develop and demonstrate an innovative integrated forest monitoring and pathway assessment system that allows reporting to European and global policies (related e.g. to greenhouse gas and forest biodiversity) and combines it with advanced forest management pathway assessment. The system starts with high-resolution mapping and precise estimation of forest attributes using a combination of field and remotely sensed data. The objective of Work Package 2 (WP2) is to develop and demonstrate a mapping and estimation platform for efficient monitoring of forest attributes including carbon stocks and biodiversity. The main outcome of WP2 is a concept for the PathFinder platform.

This deliverable 'D2.1 Initial characteristics for mapping and estimation platform', describes the initial plans on the technical implementation of the PathFinder platform. This description allows also to highlight the main characteristics of the platform and discuss potential further development options for operational implementation of such a platform in the future.

The platform aims to facilitate effective integrated use of field data with remotely sensed and other auxiliary datasets to produce precise forest information. The unique aspect of the PathFinder platform is that it enables combined use of the vast datasets of National Forest Inventories (NFIs)[1] and remote sensing data, while preserving the privacy of field plot locations. Sample plot location privacy is not only a critical precondition for the credibility of permanent sample plots but may also be subjected to privacy legislation in some countries. Utilizing these data, the platform enables various estimation approaches and models to be used to provide European wide forest information with a spatial resolution down to 10 m annually. Thus, the system facilitates reporting for the upcoming Forest Monitoring Law (EC 2023) with statistically sound methods at various levels.

The PathFinder platform consists of three main modules (Figure ES-1): 1. Data preparation module, 2. Mapping and estimation module and 3. Results module. Each of the three main modules consist of several components that together allow to conduct the tasks required in different phases of the data pre-processing, analysis and distribution. Together these three modules form a comprehensive system that allows annual processing and analysis of the Earth observation (EO), auxiliary and field datasets.

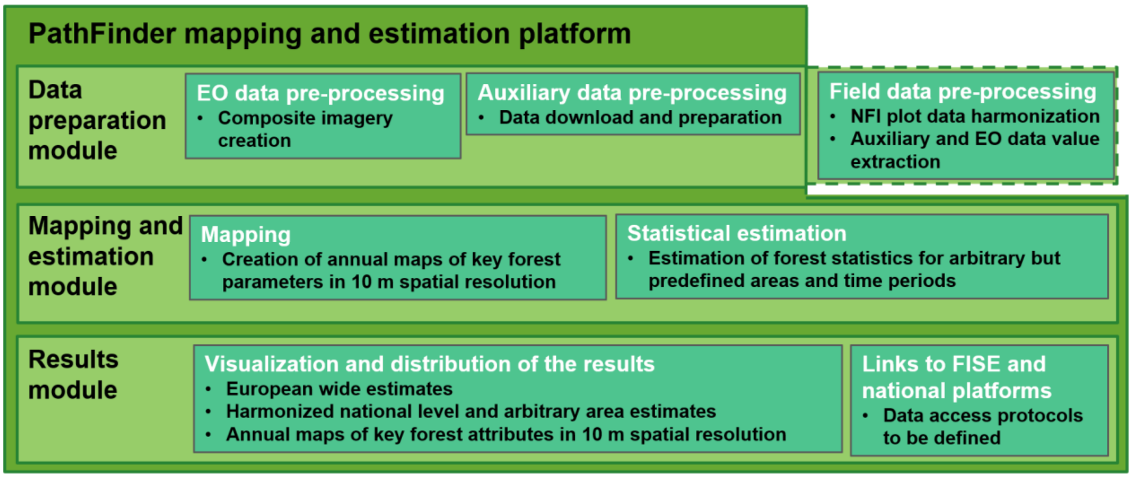

Figure ES-1. High-level description of the PathFinder mapping and estimation platform. Note that the field data pre-processing is conducted by national counterparts to maintain data privacy and ensure specific NFI-related background knowledge in data handling.

[1] Or any other field data (ground truth) obtained by probability sampling of geographical space.

All required pre-processing tasks are conducted in the ***Data preparation*** module. A large part of the field data pre-processing is conducted by national NFI teams in their own premises (due to plot location privacy reasons and the necessary, specific NFI-related background knowledge). Tools are provided to them for data processing as well as checking the correctness of the outputs. The components of the ***Mapping and estimation*** module facilitate synergistic use of EO and field datasets and thereby form the core of the PathFinder platform. This is the module where harmonized European wide forest information with a spatial resolution down to 10 m is produced annually. Statistically sound estimators enable reporting at various levels. The ***Results*** module enables visualization and distribution of the platform outputs to the users. In addition to a dedicated PathFinder portal, links to other platforms will be created to maximize the accessibility of the output products.

The *Mapping and estimation* module is the core of the PathFinder platform where the production of high-resolution maps as well as statistically sound estimates is implemented. The module closely interlinks two components:

1. Map production with a combination of EO and field data.
2. Statistically sound estimation of target parameters in predefined geographical areas (cells), eventually complementing field data by auxiliaries in the form of wall-to-wall maps.

The close linkage and combined use of the maps and estimates is one of the biggest strengths and a unique feature of the PathFinder platform concept. This allows provision of consistent time series of forest estimates while also providing the latest forest information maps. In addition to serving as input for European forest scenario modelling, the maps can also be used as inputs for improved statistical estimation (model-assisted or model-based).

The demonstration of the PathFinder platform during the project will mainly utilize Forestry TEP (https://f-tep.com/) and nFIESTA (new Forest Inventory ESTimation and Analysis, https://gitlab.com/groups/nfiesta/; Adolt et al. 2018, Kohn et al. 2018). Forestry TEP is an EO data processing platform and nFIESTA is a methodological framework and a set of software packages (PostgreSQL database extensions, GUI frontends and scripts) implementing generic and flexible estimation procedures relying on probability samples. It is important to note that PathFinder platform concept is by no means limited to using Forestry TEP or any other particular EO data processing environment.

A basis for a future forest monitoring platform is designed and demonstrated during the PathFinder project. The main goal of the demonstration is to prove the scientific and technical feasibility. One of the key areas of further development going beyond the project duration is the operational set-up of the platform. The upcoming European forest monitoring law which is currently under discussion (EC 2023) may have an impact on the management decisions of the operational system. Another important further development aspect is the expansion of EO and auxiliary datasets to retain the relevance of the system into the future as new types of datasets become available. Attention needs to be given to maintaining the consistency of the results over time when new data sources are integrated into the system. These aspects further underline the importance of the flexibility of the technical solutions chosen to be used during the PathFinder project.

V. Abbreviations

ALOS	Advanced Land Observing Satellite
ALS	Airborne Laser Scanning
API	Application Programming Interface
BFW	Federal Research and Training Center for Forests, Natural Hazards and Landscape
BOM	Byte Order Mark
CAMS	Copernicus Atmosphere Monitoring Service
CMEMS	Copernicus Marine Service
CLC	Corine Land Cover
CLMS	Copernicus Land Monitoring Service
CEMS	Copernicus Emergency Management Service
CSV	Comma-separate values (file format)
DB	Database
DIABOLO	Distributed, Integrated and Harmonised Forest Information for Bioeconomy Outlooks project
EC	European Commission
EEA	Environmental Agency
EFFIS	European Forest Fire Information System
EFMS	European Forest Monitoring System
EO	Earth Observation
EU	European Union
DIAS	Data and Information Access Services
ESA	European Space Agency
FISE	Forest Information System for Europe
FAO	Food and Agriculture Organization
FRA	Forest Resource Assessment
Forestry TEP	Forestry Thematic Exploitation Platform
GIS	Geographic Information System
GUI	Graphical User Interface
HRL	High Resolution Layers
IaaS	Infrastructure as a Service
ICESat	Ice, Cloud and land Elevation Satellite
ICT	Information and Communication Technology
INSPIRE	Infrastructure for Spatial Information in Europe
k-NN	k Nearest Neighbours
L2A	Level 2A
Lidar	Light Detection and Ranging
LUKE	Natural Resources Institute Finland
LULUCF	Land use, land-use change, and forestry
MSI	Multi-Spectral Instrument
NFI	National Forest Inventory
nFIESTA	new Forest Inventory ESTimation and Analysis
NIBIO	Norwegian Institute of Bioeconomy Research
NIR	Near Infrared
NISAR	NASA-ISRO SAR Mission
NUTS	European Nomenclature of territorial units for statistics
OGC	Open GeoSpatial Consortium
REST	Representational State Transfer
ROSE-L	Radar Observing System for Europe in L-band
SAR	Synthetic-Aperture Radar
SWIR	Short Wave Infrared
TI	Thünen Institute of Forest Ecosystems
UHUL	Czech Forest Management Institute
VTT	Technical Research Centre of Finland Ltd.

Contents

1. Introduction

The PathFinder project aims to develop and demonstrate an innovative integrated forest monitoring and pathway assessment system that allows reporting to European and global policies (related e.g. to greenhouse gas and forest biodiversity) and combines the reporting with advanced forest management pathway assessment. The system starts with high-resolution mapping and precise estimation of forest attributes using a combination of field and remotely sensed data. The precise forest information of the monitoring system feeds into a new scenario framework that forecasts future forest scenarios and outcomes of forest management alternatives, and further on to forest management pathway assessment.

One of the main targets of the project is the development of a consistent European Forest Monitoring System (EFMS) that combines field observations and remote sensing data in consistent national and EU level LULUCF Greenhous Gas Inventories and provides information on biodiversity and other key forest ecosystem services. The EFMS furthermore includes a future forest forecasting and policy pathway assessment system.

The objective of Work Package 2 (WP2) is to develop and demonstrate a platform to produce high-resolution maps and precise estimates for efficient monitoring of forest attributes including carbon stocks and biodiversity. The main outcome of WP2 is a concept of a mapping and estimation platform, the design of which is developed in the ‘Task 2.1: Developing a mapping and estimation platform’. The platform aims to facilitate effective integrated use of field data with remotely sensed and other auxiliary datasets to produce precise forest information. Key characteristics of the platform include:

1. Maximizing synergy between field and remote sensing data in provision of consistent estimates and maps
2. Preserving field plot location privacy

The platform concept is demonstrated by building on existing components (most particularly Forestry TEP and nFIESTA which are presented in Section 3). The unique aspect of the PathFinder platform is that it enables combination of the vast field datasets provided by European National Forest Inventories (NFIs) and remote sensing data. With several hundreds of thousands of NFI field plots scattered around Europe, this unprecedented volume of field data, together with the growing amounts of remote sensing data, not only provides unique possibilities but also requires new scientific and technical solutions for unleashing the full synergy of the datasets. The PathFinder platform utilizes various estimation approaches and models to provide consistent European wide forest information with a spatial resolution down to 10 m annually. Estimation approaches are implemented in nFIESTA, while high-resolution maps are produced in Forestry TEP.

As an important feature, the platform facilitates the use of field plots as reference data, while preserving the privacy of field plot locations. This approach developed earlier for estimation purposes in nFIESTA (in the context of the EU DIABOLO project[2]) is now applied also to map production. The PathFinder platform is designed in such a manner that applications requiring access to sensitive information can be run by respective authorities in their own IT environment, ensuring that data restrictions are maintained.

This deliverable ‘D2.1 Initial characteristics for mapping and estimation platform’, describes the technical implementation that will be used in the final demonstration of the platform to be conducted in ‘Task 2.4 Demonstration of the mapping and estimation platform’. This description allows us also to

[2] http://diabolo-project.eu/

highlight the main characteristics of the platform and discuss potential further development options for operational implementation of such a platform in the future.

The deliverable has been divided into six sections:

1. *Introduction*, which provides the context for the document within the project.
2. *High-level description of PathFinder mapping and estimation platform*, which describes the overall structure and key technical and scientific choices made.
3. *Forestry TEP and nFIESTA,* which present the two main technical components enabling the demonstration of the platform to be conducted during the project.
4. *Data preparation module set-up*, which describes the pre-processing components of the platform in detail.
5. *Mapping and estimation module set-up*, which describes the mapping and estimation components of the platform in detail.
6. *Results module set-up*, which describes the result visualization and distribution features in detail.
7. *Future development directions*, which discusses ideas for future development and options for operational implementation of the platform.

2. High-level description of PathFinder mapping and estimation platform,

2.1 General approach

The amount of data available for forest monitoring has been rapidly increasing over the past ten years. Particularly the amount of remotely sensed data has dramatically increased, due to the influx of 10-30 m satellite imagery (such as Sentinel 1 and 2, as well as Landsat 8 and 9) and the growing availability of < 10 m spatial resolution data and airborne lidar datasets. In combination with the National Forest Inventory (NFI) field measurements, these data enable monitoring of European forests in an unprecedented accuracy and frequency. However, to cope with the immense data quantities and various processing steps required, standardized procedures and platforms are needed.

Numerous different types of forest information tools and platforms exist in Europe, including most importantly 1) storage and processing platforms (e.g., Data and Information Access Service DIAS and Copernicus Data Space Ecosystem), 2) application platforms (e.g., Forestry TEP https://f-tep.com/), 3) estimation tools (e.g., nFIESTA; Adolt et al., 2018, Kohn et. al 2018) and 4) distribution and visualization platforms (e.g., FISE, FAO Global FRA). During the project, the PathFinder platform demonstration will mainly utilize existing components, specifically Forestry-TEP and nFIESTA, combining them with inter-platform connections and processing pipelines (Södergård et al. 2021) to take full advantage of the synergies that the existing infrastructure provides.

This diversified approach during the project provides several benefits. Firstly, no new infrastructure is needed and the most suitable infrastructure for any given component can be chosen. Secondly, by building the system from components, flexibility is retained regarding future implementation of the platform. Particularly, it is important to note that the Forestry TEP platform was chosen as the EO data processing platform as it is run by a consortium member. However, the PathFinder platform concept is by no means limited to using Forestry TEP or any other particular EO data processing environment. Different components can be moved to optimal infrastructure as appropriate. Thirdly, individual components of the platform can be extended, modified or replaced without the need of changes in the other components, as long as the input and output characteristics are maintained the same.

It is important to emphasize that during the project the main aim is to demonstrate the feasibility of the platform. This is achieved by building tools and processing components that fulfil the required functionalities of the platform and demonstrating the technical feasibility by running European wide demonstrations using the platform components. The platform functionality is demonstrated using a set of existing infrastructures as described above. In the future, in a potential operational deployment, the platform can be implemented on and operated in other processing infrastructure if so desired.

2.2 PathFinder platform structure

The PathFinder platform consists of three main modules (Figure 1):

1. Data preparation module
2. Mapping and estimation module
3. Results module

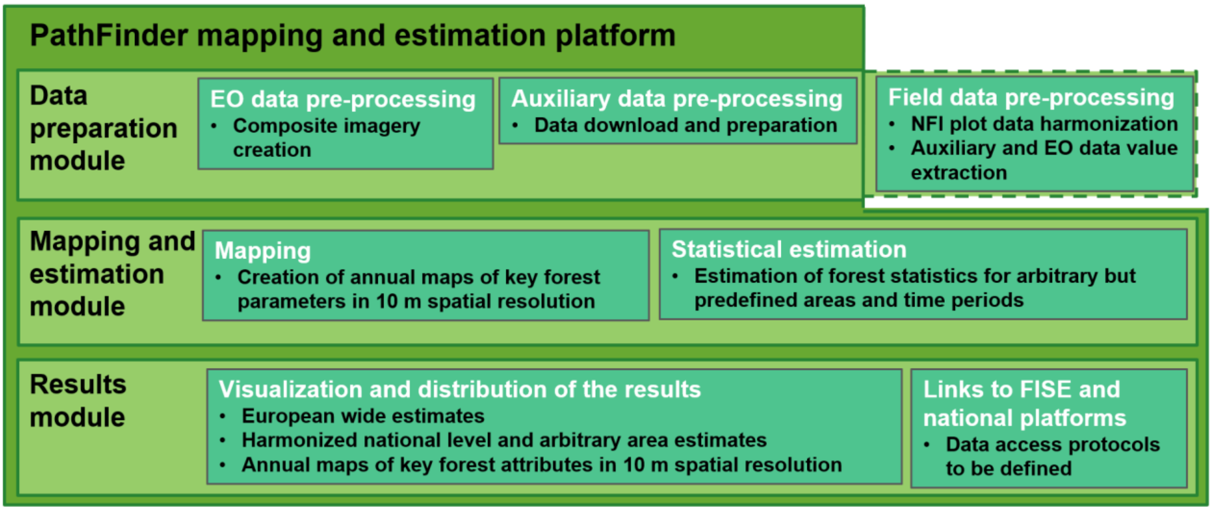

Figure 1. High-level description of the PathFinder mapping and estimation platform. Note that the field data pre-processing is conducted by national counterparts to maintain data privacy and specific NFI-related background knowledge.

Each of the three main modules consists of several components that together allow to conduct the tasks required in different phases of the data pre-processing, analysis and distribution. Together these three modules form a comprehensive system that allows annual processing and analysis of the Earth observation (EO), auxiliary and field datasets.

The work starts with the ***Data preparation*** module, which takes care of all required pre-processing tasks for EO, auxiliary and field datasets. Note that large part of the field data pre-processing is conducted by national NFI teams in their own premises (due to privacy reasons and the necessary, specific NFI-related background knowledge). Tools are provided to the NFI teams for the data processing as well as checking the correctness of the outputs. The initial *Data preparation* module set-up for the platform demonstration is described in Section 4.

The components of the ***Mapping and estimation*** module facilitate synergistic use of EO and field datasets and thereby form the core of the PathFinder platform. This is the module where harmonized forest information from 10 m spatial resolution to European level on annual basis is produced using data from the *Data preparation* module. The initial *Mapping and estimation* module set-up for the platform demonstration is described in Section 5.

The main objective of the ***Results*** module is to enable visualization and distribution of the platform outputs to the users. In addition to a dedicated PathFinder portal, links to other platforms will be created to maximize the accessibility of the output products. The initial *Results* module set-up for the platform demonstration is described in Section 6.

3. Forestry TEP and nFIESTA

The demonstration of the PathFinder platform during the project will mainly utilize Forestry-TEP and nFIESTA. Basic presentation of these two key components that form the core of the demonstration environment are provided here. However, it is important to again highlight that the PathFinder platform concept is by no means limited to using Forestry TEP or any other particular EO data processing environment.

3.1 Forestry TEP

The environment used for EO data pre-processing and map production is the Forestry Thematic Exploitation Platform (Forestry TEP, https://f-tep.com/), operated by VTT Technical Research Centre of Finland Ltd. Developed as part of ESA's EO Exploitation Platforms initiative, it is an online environment designed for the forestry community. Forestry TEP enables commercial, governmental and research users in the forestry sector globally to efficiently access satellite data-based processing services and tools for generating value-added information products, such as biomass maps or change detection products. Forestry TEP can be applied to conduct all types of forest monitoring related remote sensing-based processing and image analysis. It is continuously maintained and developed further to support the users.,

Forestry TEP operates on top of the CREODIAS (https://creodias.eu/) storage and processing platform, one of the Copernicus Data and Information Access Services (DIAS). The Forestry TEP resource configuration scales dynamically based on demand, utilizing the underlying 'Infrastructure as a Service' (IaaS) environment. Forestry TEP enables access to CREODIAS data catalogue, including Sentinel 1, 2, 3 and 5P, Landsat 5, 7 and 8, and other datasets (e.g., Copernicus core data: CAMS, CMEMS, CLMS and CEMS). The data are available as continuous global long-term coverage. Users can also upload any satellite or reference data into the platform.

Forestry TEP allows users to create and implement their own algorithms for any task (Figure 2), utilizing the datasets available on the platform or uploaded by the user. Processing services can be created in an online Linux Docker based development environment allowing utilization of numerous programming languages including e.g., C, C++, C#, Java, JavaScript, Python, GDAL, R, PHP, Fortran, Matlab. The environment also enables integration of functionalities of Orfeo ToolBox, SNAP and GDAL into user-developed processing workflows. The developed processing services can be shared to partners or customers or made available to users at large.

Forestry TEP also offers a selection of pre-prepared tools for image analysis and online access to selected image processing and GIS software. Readily available functionalities include simple value-added products such as vegetation indices, as well as computation of forest and land cover maps and change maps. Popular applications such as QGIS and SNAP can be used online on the platform through the web browser. In addition to the graphical user interface, Forestry TEP features can be accessed from external systems through Application Programming Interface (API) access points.

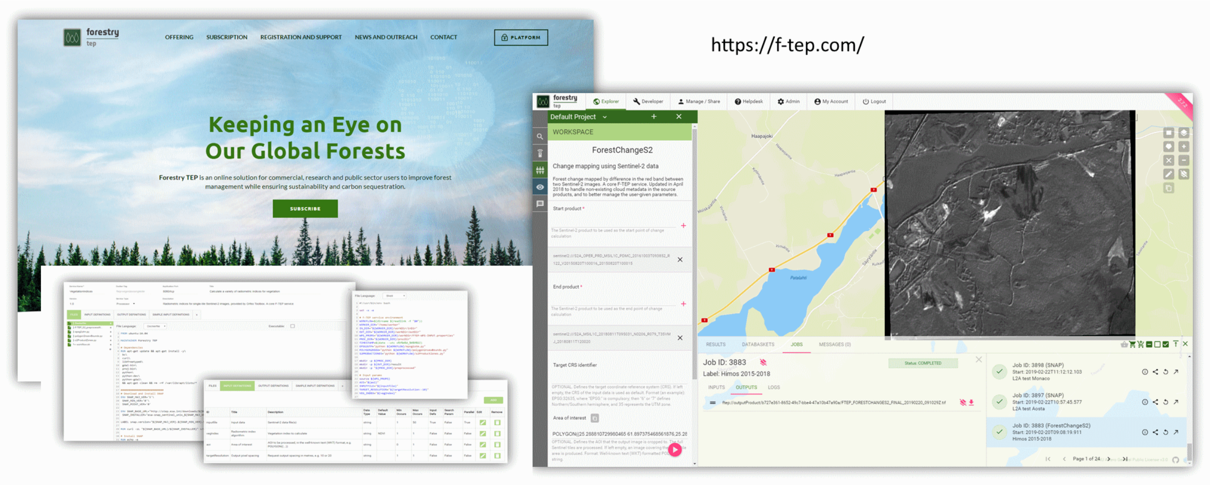

Figure 2. Forestry TEP online processing platform for forest monitoring.

During the PathFinder project, Forestry TEP is used as the main EO data processing platform. However, it is important to understand that the PathFinder platform concept is flexible in such a manner that other platforms can be used instead or in addition to Forestry TEP in the potential future operational implementation of the PathFinder platform.

The data processing operations to be conducted in Forestry TEP during the PathFinder platform demonstrations are described in detail in the following sections. Earth Observation (EO) and auxiliary data pre-processing steps are described in Section 4.2, while the mapping component operations are described in Section 5.2. Specific operations to be conducted in the mapping component include feature bank creation and map production. The direct access to EO data enables smooth implementation of the required processing steps in Forestry TEP.

3.2 nFIESTA

nFIESTA (new Forest Inventory ESTimation and Analysis, https://gitlab.com/groups/nfiesta/) is a methodological framework and a set of software packages (PostgreSQL database extensions, GUI frontends and scripts) implementing generic and flexible estimation procedures relying on probability samples of any variable(s) defined in a geographical domain (two-dimensional space).

nFIESTA is capable to produce statistically sound estimates for arbitrary geographical areas[3] from pooled probability samples. Typically, the probability samples are NFI data, which come from one or more countries, each country having a different sampling design. The geographical areas can also cross the country borders. However, there is no limitation to forest inventory data or applications. Any probability sample of any variable in a geographical area can be evaluated by nFIESTA. Field plot (e.g., forest inventory, the ground truth) and auxiliary (e.g., GIS layers produced by remote sensing) data may eventually be combined in the model-assisted or model-based setting to improve the accuracy of estimates. Doing so, the design of nFIESTA does not require disclosure of field plot coordinates. Instead, it allows users (countries, NFIs) to work with their sensitive data at their home institution and only provide the minimum necessary data to the central (integration) database, such as mappings of field plots to estimation cells and values of aggregated field data and auxiliary variables at field plot positions.

[3] Subject of provision of the plot membership to each estimation area by the countries.

The design and development of nFIESTA was formally initiated as part of the DIABOLO project[4] financed by the EU's H2020 research and innovation programme, namely within task T2.3.1. *Common data analysis and output delivery system*. A comprehensive information about the DIABOLO achievements can be found in a technical report (Ene, et al., 2018).

Beyond the DIABOLO project lifetime, the development of nFIESTA continued at the Forest Management Institute under the Czech Ministry of agriculture (ÚHÚL, www.uhul.cz). Until now, two Czech NFI cycles, NFI2 (2011-2015) and NFI3 (2016-2020), were successfully evaluated by nFIESTA.

nFIESTA is built on top of the PostgreSQL RDBMS (https://www.postgresql.org/). Its modules use PostgreSQL extensions (https://wiki.postgresql.org/wiki/Extensions) as database and functional backends (API).

Conditions of use and further development on top of the existing nFIESTA code are regulated by the EUPL. nFIESTA is free and open source. Since the end of 2018, the nFIESTA GUI (Home · Wiki · nFIESTA / nfiesta_gui · GitLab) has been developed and more functionality was added to the prototype that was originally aimed at generic estimation only. Currently implemented nFIESTA functionality (as to the beginning of 2024) is described in Figure 3 (next page).

The heart (statistical estimation) of nFIESTA can be seen in the middle, pink rectangle (Figure 3), entitled Central nFIESTA database. Target parameters can be statistically estimated by directly calling the functions included in the core PostgreSQL extension *nfiesta_pg* (the white DB icon in the smaller green rectangle, Home · Wiki · nFIESTA / nfiesta_pg · GitLab). However, it is much more convenient to use the nFIESTA-GUI module *'Estimates'* that can be run on a Windows PC and connect remotely to any version-compatible instance of *nfiesta_pg*. The currently implemented estimators were described in a technical report of DIABOLO (Adolt, Lanz, & Fejfar, 2018).

The *nfiesta_pg* uses information about the sampling design attached to any (NFI) data it processes. This information is managed within the *nfiesta_sdesign* extension (the white DB icon in the two yellow rectangles, Home · Wiki · nFIESTA / nfiesta_sdesign · GitLab). There is also the *nfiesta_gisdata* PostgreSQL extension (inside the grey rectangle, Home · Wiki · nFIESTA / nfiesta_gisdata · GitLab) that, together with the GUI module 'GIS data' can be used for auxiliary data pre-processing.

There are two options how NFI data can be uploaded to the central nFIESTA database. The default option (grey rectangle on the right-hand side of the diagram) is to upload data in the form of CSV files created by NFIs following a standard specification. This option is being used for the ongoing PathFinder demonstration study and it was used in DIABOLO case study for the first time (Lanz, et al., 2018).

A comprehensive description of the CSV files, their structure and quality control procedures (note the turquoise vertical strip in the diagram) can be found on the Pathfinder demonstration Wiki Description of CSV files · Wiki · nFIESTA / pathfinder_demo_study · GitLab). The description is based on the original CSV specification elaborated within DIABOLO (Lanz, Adolt, Kohn, & Fejfar, 2018) that has recently been updated to incorporate the past DIABOLO development.

The left-hand side of the diagram (the big pale green rectangle) describes the second option how NFI and auxiliary data can be uploaded to the central nFIESTA database. In this (preferred) scenario the plot level field data as well as auxiliary data are prepared by using nFIESTA extensions and GUI modules.

[4] http://diabolo-project.eu/

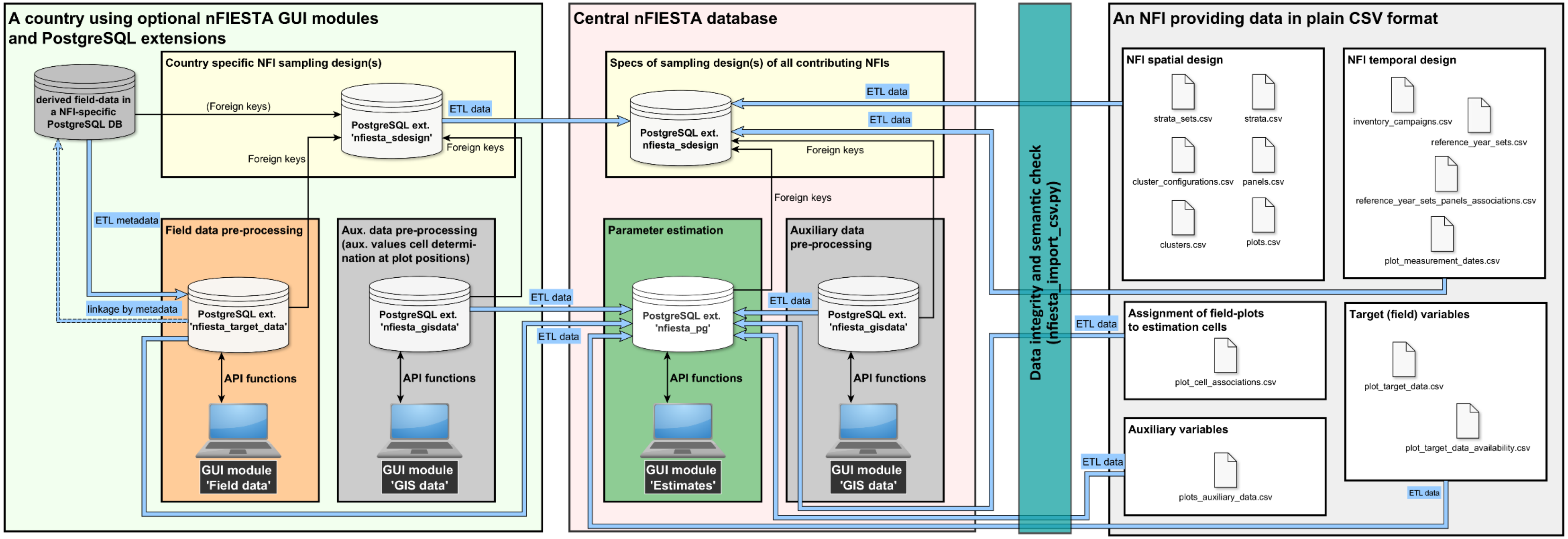

Figure 3. Diagram of the main functional components of nFIESTA.

The *nfiesta_gisdata* extension and GUI module 'GIS data' have already been mentioned. The PostgreSQL extension *nfiesta_target_data* (Home · Wiki · nFIESTA / nfiesta_target_data · GitLab) and the GUI module 'Field data' (both inside the smaller, orange rectangle) serve the preparation of plot-level values of target variables, the so-called local densities, by aggregating contributions (numerical values) of specific objects, e.g., trees or dead wood pieces. The aggregations typically involve categorisation by attributes of the objects themselves, e.g., species, and/or by attributes of other objects lying higher in the NFI data hierarchy, e.g., altitude attached to the plot centre[5]. The resulting local densities are saved in the local database hosting the *nfiesta_target_data* extension. At any time, they can be transferred (by using a set of PostgreSQL API functions and GUI, or via CSV export) to a target database hosting the *nfiesta_pg* extension, where the statistical estimation can proceed. The target DB can be the country's local nFIESTA database or the central database integrating data from more countries.

The extension *nfiesta_target_data* needs to be connected to 'derived' NFI database (the dark grey DB stack in the upper left corner of the pink rectangle) that complements the 'raw' NFI data by derived attributes such as tree volumes divided by inclusion zone areas, predicted tree heights, results of soil sample analysis and similar. The structure of the derived NFI database is made accessible to *nfiesta_target_data* by listing object types (database tables), their hierarchies (relational structure) and attributes (columns of given tables) that can take a role of local density contributions, categorisation attributes or both. The 'derived' database is linked to the *nfiesta_sdesign* extension describing the country specific NFI design. The information about data availability in specific inventory campaigns and for representative sets of plots (panels) is also contained in the 'derived' database and it is used in the process of local density aggregations at the plot-level.

There are more extensions (modules) of nFIESTA, the complete list is included in the main GitLab page here: https://gitlab.com/nfiesta. Worth mentioning are at least the *nfiesta_htc* (nFIESTA / nfiesta_htc · GitLab) and *nfiesta_results* (Home · Wiki · nFIESTA / nfiesta_results · GitLab) extensions. The first is very fundamental because it implements the Cordy's Horvitz-Thompson theorem for continuous (infinite) populations (HTC) that is behind any statistical estimation implemented by nFIESTA. This extension is called from within *nfiesta_pg* and has not changed much since its first implementation in DIABOLO. The purpose of the second extension (*nfiesta_results*) is to serve as a DB back-end for lookup and presentation of nFIESTA estimates, and it can be used, e.g., by a web service.

4. Data preparation module set-up

4.1 Module overview

The objective of the *Data preparation* module is to pre-process all the datasets required in the subsequent processing in such a manner that it ensures seamless entry of datasets into the *Mapping and estimation* module. The module has two main components:

1. EO and auxiliary data pre-processing
2. Field data pre-processing

[5] Imagine, e.g., growing stock as attribute of single trees that needs to be aggregated at plot level by categories of species (recorded at the single tree level) and by categories of altitude above sea level (recorded at the plot level, thus common for all trees on a plot).

In both components, the implemented tools and processes take the raw inputs, do necessary calculations and conversions, and save them into the formats required by further processing in the platform. The EO data pre-processing is conducted in the Forestry TEP platform in the demonstration, but it can be conducted in any EO data processing environment. The field and auxiliary data pre-processing is done by national NFI teams in their own premises (due to privacy reasons and the necessary, specific NFI-related background knowledge) with the final EO and auxiliary layers provided through Forestry TEP. As outputs, the *Data preparation* module provides 1) the final EO and auxiliary layers and 2) the NFI (harmonised) data in nFIESTA format, which in turn serve as input for the *Mapping and estimation* module.

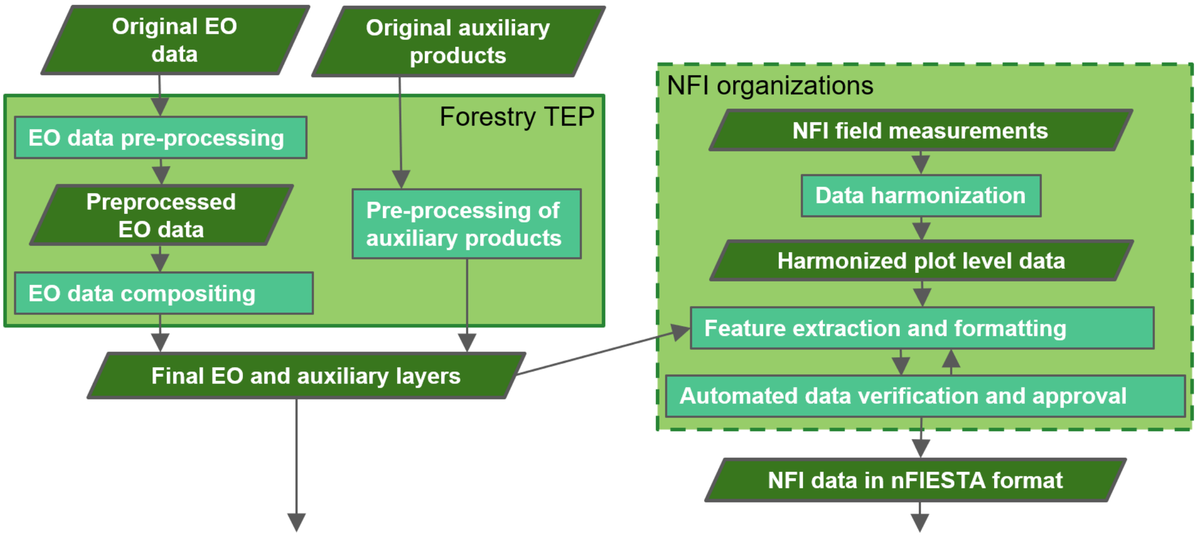

Figure 4. Overall workflow of the Data preparation module of the PathFinder mapping and estimation platform.

In future applications, field and auxiliary data pre-processing could be performed by using the nFIESTA GUI modules 'Field data' and 'GIS data' and the corresponding database extensions, i.e., bypassing the CSV files, see description in section 3.2 and Figure 3. Once these nFIESTA modules are fully developed, the field and auxiliary data quality can be better guaranteed due to more sophisticated checks and a seamless transfer of data to the nFIESTA central database (*nfiesta_pg* extension). In addition, every step of field data processing by the nFIESTA GUI modules is recorded in a JSON metadata file (for example the use of specific models to predict biomass of trees, or the general DBH threshold) that accompanies the NFI data until the lookup and presentation of statistical estimates produced by nFIESTA.

4.2 Earth Observation (EO) and auxiliary data pre-processing

Pre-processing of EO data: In general terms, the Earth Observation (EO) data pre-processing needs to be able to take care of all the pre-processing tasks of EO and auxiliary datasets. Depending on the type of datasets used, these pre-processing steps may include geometric or spectral corrections, resampling, reprojection, mosaicking, and compositing, among other tasks. Due to the European wide focus of the platform, all the data pre-processing steps need to be conducted for high volumes of data. In an operation setting, an annual processing cycle is foreseen.

Currently, the most important Earth Observation dataset for the EFMS are optical Sentinel 2 data from the European Space Agency. With the two satellites in orbit, it provides 2-3 days revisit time in Europe. The Multi-Spectral Instrument (MSI) on board Sentinel-2 satellites has 13 spectral bands, four of which have 10 m and six of which have 20 m spatial resolution. The remaining three bands with 60 m spatial resolution are mainly used for atmospheric correction. The high imaging frequency allows the production of growing season cloud free composite images.

The PathFinder platform demonstration is conducted using Sentinel-2 cloud free composite imagery data as the EO dataset (see more detailed description of the processing below). Sentinel-2 was selected as the EO dataset due to its high temporal and spatial resolution. Furthermore, the continuity of Sentinel-2 type data is ensured by the Copernicus space programme as they form the basis for many of the Copernicus products and services[6]. In the future, inclusion of other types of EO data can be considered. Please see more discussion on this topic in Section 7.

Pre-processing of Auxiliary data: The EO data can be supported by auxiliary datasets (such as the Copernicus products). For the European forest monitoring purposes, these include most particularly the Corine Land Cover map (CLC, and in the future its new versions CLC+), and the pan-European forest High Resolution Layers (HRL), which are foreseen to be annual products in the future. Supporting information may be available also from the European Forest Fire Information System (EFFIS), and other similar systems providing information on natural and anthropogenic events occurring in the forests.

From the point of view of statistical estimation, the existing auxiliary datasets can improve precision of target parameter estimates, which will be particularly useful in situations of limited amount of field data. In addition, and under the precondition that a good model linking field and auxiliaries can be fitted, model-based estimates can be produced for geographical areas and time point where no field data is available at all[7]. nFIESTA implements pre-processing of auxiliary data in one of its PostgreSQL extensions, *nfiesta_gisdata* (Home · Wiki · nFIESTA / nfiesta_gisdata · GitLab), and GUI modules ('GIS data'), see the grey rectangle inside the diagram in Figure 3. Figure 5 shows a screenshot of the 'GIS data' GUI module. The user instructions for this module can be found on a wiki page here: GIS data module · Wiki · nFIESTA / nfiesta_gui · GitLab

Three following tasks are currently implemented:

1. Assign plots to estimation cells (geographical areas for which estimates need to be calculated). This is done by intersections of the plot coordinates with the geometries of estimation cells in the PostgreSQL DB (PostGIS vector format).
2. Determine the value of an auxiliary variable for an exact plot position by spatial intersection with a GIS layer (stored in the PostgreSQL database as PostGIS vector or raster).
3. Evaluate the total of auxiliary variables for a collection of cells.

The first two tasks require access to precise plot coordinates, so they are typically not done at the central nFIESTA database. This is because the provision of precise NFI coordinates has rather been an exception so far and there are no plans to change this in the foreseeable future. The reason for this is to prevent any actual or seeming loss of representativeness of the NFI sample plots. Under the assumption that auxiliary values were not manipulated by adding random noise or by a functional transformation (to prevent from indirect disclosure of plot coordinates), the third task can be solved at the central level. It is however more likely, that the 'GIS data' module is used by individual countries and only the resulting data is provided to the central database.

[6] https://www.copernicus.eu/en/copernicus-services

[7] Note, that ALL nFIESTA estimates are ALWAYS accompanied by accuracy measures and metadata based on which the quality of each single estimate can be assessed. At least from the perspective of nFIESTA use-case, model-based estimates are considered qualitatively inferior compared to design-based as well as model-assisted estimates. The reason is that the accuracy of model-based estimates is conditioned by model validity, for more details, see section 5.3.2.

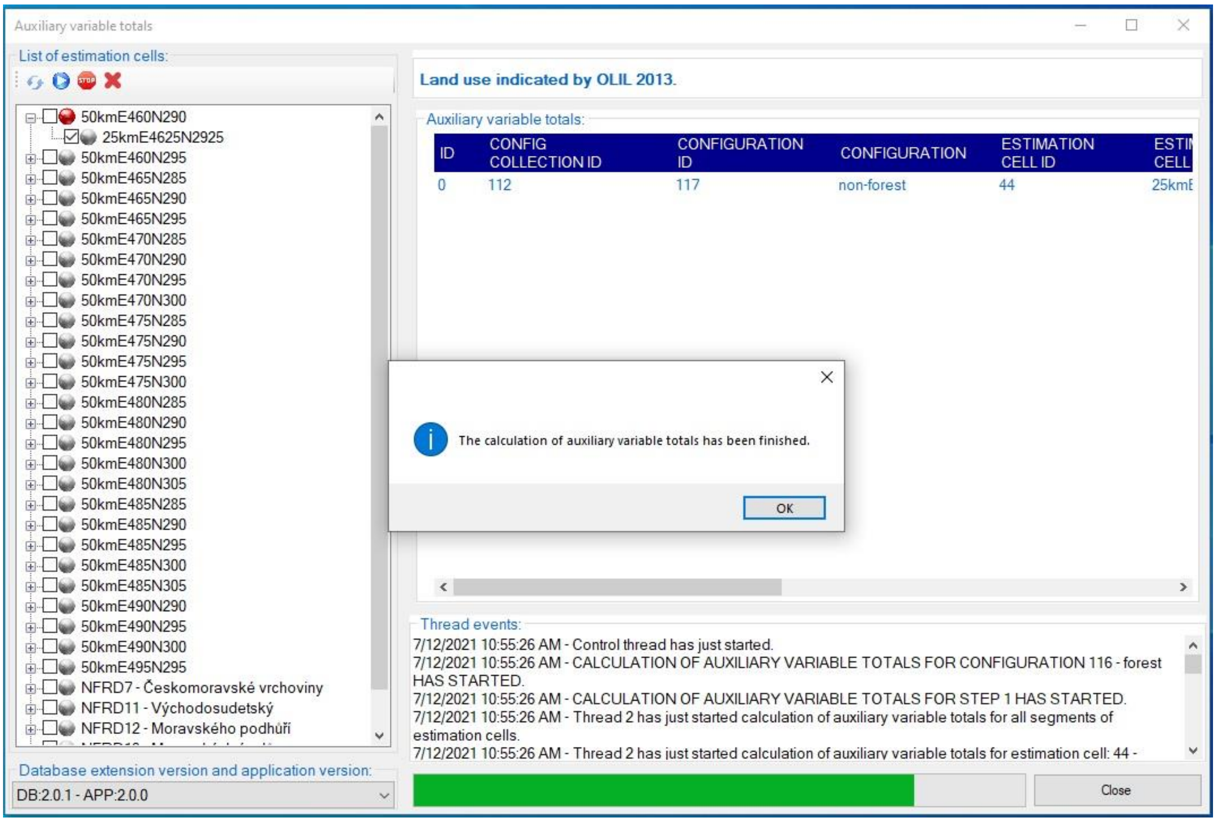

Figure 5: Screenshot of nFIESTA GUI module 'GIS data' used for calculation of auxiliary variable totals by estimation cells.

Alternatively, countries may prepare the corresponding data by their own means instead using the CSV files modality, as described in Section 4.3.1.

The demonstration of the mapping and estimation platform (project deliverable D2.4) will explore the potential for improvement of the estimation accuracy of forest area, above ground biomass, harmonised stem volume, and changes of these parameters by incorporation of auxiliaries in addition to NFI field data. The following three auxiliary layers are used in the demonstration:

1. Copernicus Land Monitoring Service (CLMS) Tree Cover Density[8] 2018
2. CLMS Forest Type[9] 2018
3. Global Forest Change[10] (Hansen et al. 2013)

Values of these layers were already determined at the NFI plot coordinates of the involved countries and uploaded to the central database, see Auxiliary variables · Wiki · nFIESTA / pathfinder_demo_study · GitLab. It is important to highlight that the platform concept is not limited to using the three auxiliaries listed above but can accommodate any auxiliary datasets. Issues related to the independence of the auxiliary layers should be considered when selecting the auxiliaries.

Preserving plot location privacy: A special aspect of the PathFinder approach is that all the EO and auxiliary layers that will be used in the mapping and estimation phase are transferred to the national counterparts who take care of the extraction of the EO and auxiliary data values for each field plot. This way the plot location information can be retained within the institutions.

[8] https://land.copernicus.eu/en/products/high-resolution-layer-tree-cover-density
[9] https://land.copernicus.eu/en/products/high-resolution-layer-forest-type
[10] https://glad.earthengine.app/view/global-forest-change

In the derivation of a feature bank (i.e., a combination of field, EO and auxiliary data values) the pixel values are calculated as weighted averages by plot coverage. However, due to the approximate plot location (e.g., linkage to 1 km x 1 km INSPIRE cells) also provided elsewhere in the database, by combining these data it is possible to reveal the actual plot position. Further steps (e.g., adding noise to the spectral values or removal of the approximate location) may be required to keep the plot location concealed in an operational setting in the future. This topic will be studied during the project with the demonstration feature bank to find an optimal solution.

Platform agnostic approach: A key feature of the Earth Observation (EO) and auxiliary data pre-processing component is the flexibility to use different platforms or existing high-level products as long as the outputs follow the Open GeoSpatial Consortium (OGC) standards (https://www.ogc.org/standards/). The EO and auxiliary data pre-processing can be performed on any or several platforms, or analysis-ready-products can be used. The only requirement is that the final pre-processed data layers need to be available for the national counterparts conducting the feature extraction and later on for the *Mapping and estimation* module.

Demonstration study pre-processing: For the PathFinder platform demonstration, the Earth Observation (EO) and auxiliary data pre-processing is conducted on the Forestry TEP platform. Annual growing season cloud free Sentinel-2 composite images are used as the EO dataset. The feature bank is created with 2020 and 2021 satellite data. The compositing was conducted with an approach developed by Terramonitor[11] and described in detail at Miettinen et al. (2021). The processing was done in Sentinel-2 tiles, which is the standard format in which Sentinel-2 data is made available.

To build snow-free growing season composite images, imagery between 15th of May and 15th of September were used, except for area north of 60°, where only images between 15th of June and 15th of September were accepted. The final composite image pixels were weighted averages of the available cloud free observations (Miettinen et al. 2021). The final composite images included seven spectral bands (B02 Blue 0.49 µm central wavelength, B03 Green 0.56 µm, B04 Red 0.67 µm, B05 Red Edge 1 0.71 µm, B08 NIR 0.84 µm, B11 SWIR 1.61 µm and SWIR 2.19 µm) and three quality bands. These bands were selected based on earlier results on optimal set of bands for forest variable prediction (Astola et al. 2019, Miettinen et al. 2021). All bands were resampled to 10 m spatial resolution using the nearest neighbour resampling.

The auxiliary datasets (Table 1) were used in their original format. Both the pre-processed EO datasets as well as the auxiliary layers were subsequently placed on a sftp portal, from where the national counterparts downloaded the datasets to their system for feature extraction. All of the layers were also retained in the Forestry TEP platform for further use in the *Mapping and estimation* module (Section 5).

4.3 Field data pre-processing

4.3.1 Preparation of plot level values

Field data entering the estimation and mapping module of the platform comes from probability samples of two-dimensional space, see Section 5.3 and the technical reports reference therein. The samples are generated by sampling designs and according to specific temporal plans. The sampling designs as well as the temporal plans (both together are further referred as survey designs) are specific for each NFI (country). In addition, the survey designs unavoidably evolve in time in response to changing information needs and NFI implementation context (budget, human and other resources,

[11] https://www.terramonitor.com/

technological development and so on). Sooner or later each NFI develops its own and specific data processing architecture tailored to its survey design and approach to statistical evaluation. Consequently, the raw NFI data differ a lot from country to country and cannot be processed without a transformation to a common, generic standard.

Such a standard was first formulated in DIABOLO project (Lanz, Adolt, Kohn, & Fejfar, 2018). At that time, the standard had the form of specification of CSV files containing the plot level values of target variables and the respective NFI survey design metadata. Currently, all data uploaded or processed by nFIESTA need to be linked to description of the NFI survey design. This is stored and managed by the PostgreSQL extension *nfiesta_sdesign* (Home · Wiki · nFIESTA / nfiesta_sdesign · GitLab, also see the two light-yellow rectangles within the diagram in Figure 3).

There are two options of NFI data upload to *nfiesta_pg* (Home · Wiki · nFIESTA / nfiesta_pg · GitLab) extension for later estimation of target parameters:

- Upload of CSV files (see right-hand side of diagram in Figure 3)
- Transfer of NFI plot-level data prepared by the *nfiesta_target_data* extension and GUI module 'Field data'

CSV files were chosen for the PathFinder demonstration study to avoid the overhead of installation and training of each NFI team on the use of nFIESTA's GUI module 'Field data', and also to avoid the need of linking the module to each NFI's specific data infrastructure (note the linkage between *nfiesta_target_data* extension and derived NFI field data in Figure 3).

The CSV files need to be prepared by each data provider (NFI or any other data representing a probability sample of geographical area) according to a standard specification. For the PathFinder demonstration study the specification has been drafted on a wiki page here: Description of CSV files · Wiki · nFIESTA / pathfinder_demo_study · GitLab. Only the structure and semantic of the data in the CSV files, but not the way how CSVs are technically created, has to be standardised. Countries use their existing or preferred approach.

The second option assumes preparation of plot-level data in a standardised way by using *nfiesta_target_data* (Home · Wiki · nFIESTA / nfiesta_target_data · GitLab). The main steps of the local density evaluation at the plot level are the following:

- Choice of an existing or definition of a new group of local density objects
- Choice of an existing or definition of a new target variable
- Categorisation of areal local density by subpopulations (optional)
- Categorisation of local density by subpopulations or area domains (optional)
- Calculation of plot-level local densities

Each of the steps is described by corresponding use-cases on the nFIESTA wiki page here: Evaluation of local densities at the NFI plot level · Wiki · nFIESTA / nfiesta_target_data · GitLab. The use-cases are implemented by the GUI module 'Field data'. Figure 6 shows one step in the selection/definition of target variable 'above ground biomass' with its associated metadata.

The prepared plot-level data, so-called local densities, need to be uploaded to the *nfiesta_pg* database extension (local or central), where statistical estimates can be configured and calculated at any time after the upload. The upload process itself is described on a wiki page here: Upload of calculated local densities to nfiesta_pg for estimation · Wiki · nFIESTA / nfiesta_target_data · GitLab.

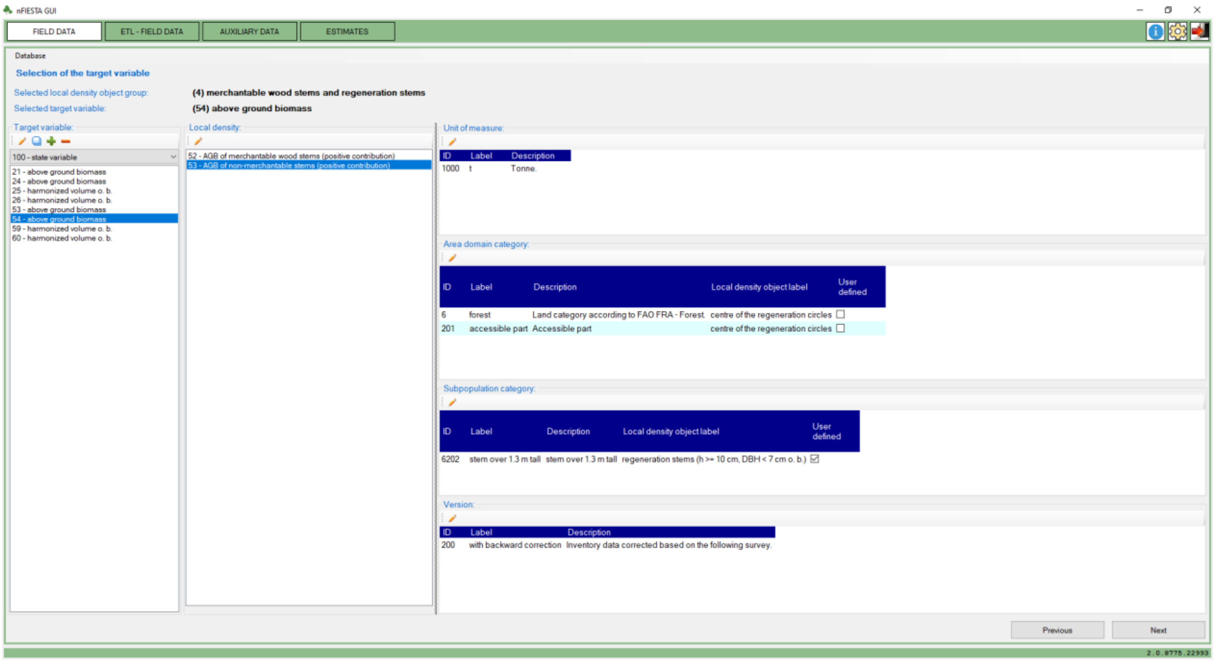

Figure 6: Screenshot of the nFIESTA module 'Field data', definition or selection of a target variable.

In case the local densities of a specific NFI are uploaded for the first time, for a new inventory campaign or for an additional panel and reference-year set combination, it is necessary to upload the NFI-specific metadata first. This procedure is formally described here: NFI survey design upload and synchronisation · Wiki · nFIESTA / nfiesta_target_data · GitLab.

4.3.2 Field data entry and Quality Control

The quality of NFI data is checked in several stages of its processing by nFIESTA (when using *nfiesta_target data* extension and the GUI module 'Field Data') and always on upload to the *nfiesta_pg* database for later estimation.

If the NFI data enter to *nfiesta_pg* in the form of CSV files, the following checks are conducted by the nfiesta_import_csv.py (Python program, see the turquoise vertical strip in Figure 3):

- Completeness of a list of CSV files.
- Compliance of each CSV file with the specified encoding, separator, quotes, end of lines, NULL values and BOM inclusion/exclusion (requirements common for all CSV files).
- Presence and correct naming of all columns (requirements specific for each of the CSVs).
- Integrity of identifier values (compound or natural foreign keys) between various CSV files making the correct linkage, e. g., between NFI data and NFI design possible.
- Check if NFI-provided statistical sampling weights sum up to the total area of sampling frame[12]
- Check if cells (geographical areas) to which sample plots are mapped at least touch the geometry of the sampling strata.
- Check of additivity of plot-level values, i.e., if sums of values (local densities) for particular categories sum up to the aggregated value for the whole plot, or for a superordinate category according to a predefined hierarchy (e.g., if sum of biomass for all species must match the value of biomass for the whole plot).

[12] The whole geographical area that is covered by a representative sample of plots (designs using single plots) or that is covered by anchor points of plot clusters (designs using plot clusters).

The instructions for the performance of CSV file checks can be found on a wiki page here: CSV files quality control · Wiki · nFIESTA / pathfinder_demo_study · GitLab.

Despite the above checks semantic errors can occur, typically because of NFI survey design misunderstanding and erroneous description in the CSV files. The delivered plot values of target variables may contain errors because they had to be extracted from the NFIs´ internal system, which may not be a standard task, so an ad hoc solution had to be created. This type of errors can be detected by uploading the prepared CSVs into a local installation of nFIESTA, configuring the basic estimates (without any auxiliaries) and comparing their values to the estimates produced by the national system. Docker containers with preinstalled nFIESTA are distributed to NFIs together with the nFIESTA GUI modules, so the comparison of nFIESTA and national estimates can be done before CSVs are sent to the central level.

If NFI data is prepared by the *nfiesta_target_data* extension and the GUI module 'Field data', most of the checks needed for CSV files do not make sense. The data are processed inside a relational database with permanent integrity checks (foreign keys) among tables of *nfiesta_target_data* and also *nfiesta_sdesign*, with primary and unique keys, not null and similar constraints, so the risk of breaking data integrity is minimised. Checks of additivity of (categorised) plot values are performed twice:

1. Immediately after calculation of plot-level local densities, but before saving them to the corresponding database tables;
2. On upload to the *nfiesta_pg* extension, before the values are saved in this extension.

The transfer of plot level local densities between the local *nfiesta_target_data* and central *nfiesta_pg* is implemented in the GUI. A direct connection over the internet (eventually using secured protocol) between the source and target DB is established. All metadata is synchronised and the transfer configuration is saved into a log. The next transfers can use this configuration and proceed faster and without the need of repeated pairing of variables and categorisations.

In the long term, it is recommended that NFIs providing data to the central database prepare their datasets using nFIESTA. In addition to the minimisation of data integrity issues, the local densities are accompanied by standardised metadata describing all important aspects of their preparation. This metadata is inherited between the phases of NFI data processing. It is envisioned that it will be useful to allow users to look up the desired estimates stored in the *nfiesta_results* (see, Home · Wiki · nFIESTA / nfiesta_results · GitLab).

4.4 Input datasets for platform demonstration

4.4.1 EO and auxiliary datasets

Input datasets used in the PathFinder platform demonstration are listed in Table 1. The auxiliary datasets are used in their original format in the feature extraction, while the Sentinel-2 L2A products are used as input for production of cloud free growing season composite images for 2020 and 2021, as described above.

Table 1. Input EO and auxiliary dataset used in the PathFinder platform demonstrations.

Dataset	Type	Characteristics
Sentinel-2 L2A	Optical satellite imagery	European Space Agency satellite, Level 2 surface reflectance, 2020 and 2021
Tree Cover Density	EO based product	Copernicus Land Monitoring Service (CLMS) Tree Cover Density[13] 2018
Forest Type	EO based product	Copernicus Land Monitoring Service (CLMS) Forest Type[14] 2018
Forest Change	EO based product	University of Maryland Global Forest Change[15] (Hansen et al. 2013) v1.9

4.4.2 Field datasets

At least the NFI data listed in Table 2 will be used in the final demonstration as it has already been collected and prepared. Additional variables may still be included. Further discussion on the additional variables can be found in the text below. The above target variables have been provided by the majority of the project partners already. The summary of the datasets gathered up to moment of this report compilation is found in Table 3.

The column '*Sample size (# of clusters or single plots)*' was calculated as follows. Each cluster or single plot (depending on the design of the NFI) was multiplied by the number of time points (surveys) in which it was assessed. These products where finally summed up. So, in contrast to the prevailing but less informative approach, the reported sample size does not refer to the sample of the geographical space only (two-dimensional space), but also the time axes (three-dimensional space). The column '*Number of plots times the number of their inventories*' was derived analogously.

The column *'Mean sampling weight of the last NFI dataset [ha year]'* shows the mean number of hectares represented by a plot or cluster in a single year of the last NFI campaign that was included in the dataset. It is an average sampling weight, knowing that some NFIs use some sort of importance sampling with unequal plot or cluster weights.

The last column in the table *'Mean sampling weight in the period of the whole dataset [ha year]'* is an analogy to the previously mentioned, but this time the mean sampling weight is calculated for the whole period between the first and last data provided (between '*Data begin [year]*' and '*Data end [year]*'). So, if a country does not have a continuous NFI, the breaks in between the NFI cycles increase the mean sampling weight expressed in ha × year units.

It is expected that the number of NFI datasets will still increase by about 7-10 countries. In addition, the list of variables has grown recently as it can be seen in the PathFinder demonstration wiki page here: Codelist of target variables · Wiki · nFIESTA / pathfinder_demo_study · GitLab. However, not all the variables in the extended set will necessarily be included in the platform demonstration. With the additional variables, it will be also possible to estimate and map indicators related to biodiversity. The final list of variables to be included in the platform demonstration will be defined on course of the project.

[13] https://land.copernicus.eu/en/products/high-resolution-layer-tree-cover-density
[14] https://land.copernicus.eu/en/products/high-resolution-layer-forest-type
[15] https://glad.earthengine.app/view/global-forest-change

Table 2. Target variables obtained by NFI field surveys that will be used in the final demonstration. For potential additional variables, please see discussion in the text.

Target variable	Definition
forest [dimensionless]	Indicator value coding the plot membership to the forest land category according to the FAO definition. In case the FAO definition cannot be met, the national definition closest to the FAO definition can be used. Please indicate which definition applies in a data readme.txt file that is submitted with the csv files. The only allowed values are 1 (the plot belongs to the forest land category) and 0 (the plot is outside the forest land category). In the ideal and simplest case, the value is determined based on the situation on the plot centre.
vol [m^3ha^{-1}]	Plot-level volume (growing stock) according to the definition No. 2 in Table 2 of (Gschwantner et al (2019): Harmonisation of stem volume estimates in European National Forest Inventories, Annals of Forest Science, 76:24, Page 6; https://doi.org/10.1007/s13595-019-0800-8**).** Non-zero values are allowed also for plots (with centres) outside the *domain_vol* if the doman is still intersected by the plot. Growing stock is the above ground volume of living stems above stump over a specified area. Included is the stem volume from the stump height to and including the stem top and the bark. Branches are excluded. Trees below 1.3 m in height, shrubs, dead trees and lying trees are not included in the selected volume definition (COST E43).
domain_vol [dimensionless]	This variable indicates whether the plot centre belongs (value set to 1) or does not belong (value set to 0) to the domain in which the target variable *vol* is defined. The domain usually comprises all forest plots, but there are exceptions in some countries: Austria, for instance, defines volume on the domain 'productive forest' (excluding for instance forest roads, protective forest without yield, inaccessible forest). In this case the *domain_vol* is 1 for all plots falling into 'productive forests' and 0 for all other plots (non-forest plots, OWL plots, non-productive forest plot etc.). The plot centre decides whether the plot belongs to the *domain_vol*. Therefore, set the value to 0 for 'shared plots' with non-zero *vol* but centre outside the *domain_vol*.
agb [tha^{-1}]	Plot-level above ground biomass following the definition SC13/17. Non-zero values are allowed also for plots (with plot centres) outside the *domain_agb* if the domain is still intersected by the plot. Harmonized AGB definition includes following components: 1. Above ground part of stump, 2. Stem from stump to top of the tree (0 threshold for both top diameter and for trees to be included), 3. Dead branches, 4. Living branches, 5. Foliage. Thus, the AGB definition includes all the components of a tree above the ground level, including stump, but not the part of stump that is below ground. Trees below 1.3 m in height, shrubs, dead trees and lying trees are not included in the selected AGB definition (SC13/17).
domain_agb [dimensionless]	This variable indicates whether the plot centre belongs (value set to 1) or does not belong (value set to 0) to the domain in which the target variable *agb* is defined. The plot centre decides whether the plot belongs to the *domain_agb*. Therefore, set the value to 0 for 'shared plots' with non-zero *agb* but plot centre outside the *domain_agb*. In most cases the *domain_agb* and *domain_vol* will be identical. Two domain variables are added for sake of robustness.
vol_change [$m^3ha^{-1}year^{-1}$]	Plot-level change of the volume (growing stock) according to the above-mentioned definition. Values correspond to the average yearly difference between two inventories conducted on the particular plot. Positive values correspond to an increase, negative to a decrease and zero to no-change situation.
agb_change [$tha^{-1}year^{-1}$]	Plot-level change of the above ground biomass following the above- mentioned definition. Values correspond to the average yearly difference between two inventories conducted on the particular plot. Positive values correspond to an increase, negative to a decrease and zero to no-change situation.

Table 3. Summary of NFI field data provided for the first phase of PathFinder platform demonstration.

Country	Number NFI cycles in the dataset	Data begin [year]	Data end [year]	Cluster (c) or single plot (s) NFI design	Number of plots times the number of their inventories	Sample size (# of clusters or single plots)	Mean sampling weight of the last NFI dataset [ha year]	Mean sampling weight in the period of the whole dataset [ha year]
AT	1	2016	2021	c	22328	5582	9020	9020
BE	3	1994	2022	s	64216	64216	1000	760
CH	3	2004	2022	s	52743	52743	1800	1490
CZ	2	2011	2020	s	59134	59134	1000	1330
DE	2	2011	2017	c	44730	11238	6350	22220
FI	1	2019	2022	c	62440	7192	19260	19260
FR	1	2017	2021	s	180920	180920	1520	1520
IE	4	2004	2022	s	69692	69692	400	1900
IT	1	2017	2020	s	6993	6993	4310	17240
NO	3	2011	2021	s	48406	48406	7360	7360
PL	3	2011	2022	c	237239	47559	7890	7890
SE	2	2012	2021	c	83960	11526	32820	41210
SI	2	2012	2018	s	2536	2536	1600	5600

Finally, it must be noted, that not all NFI data are strictly obtained in the field. Several NFIs use information from other sources such as past inventories and high-resolution aerial or satellite imagery to decide whether a sample plot (or cluster) will be visited in the field or not. Plots that are clearly outside forest or any other target land category are not visited in the field. So, the non-forest plots relatively often lack true field verification. However, the forest plots always use to be surveyed or at least verified (inaccessible plots) in the field. The potential error resulting from this procedure is usually neglected by NFIs because the error's expectation is very close to zero.

4.5 Data preparation module output datasets for platform demonstration

The output datasets from the *Data preparation* module for the platform demonstration are listed in Table 4. These include three different types of products, namely EO datasets, auxiliary datasets and nFIESTA input files. Detailed description of the various nFIESTA .csv input files is provided at: https://gitlab.com/nfiesta/pathfinder_demo_study/-/wikis/Description%20of%20CSV%20files

Table 4. Data preparation module output datasets.

Dataset	Type	Features
Sentinel-2 composite imagery 2020	EO (.tif)	B02, B03, B04, B05, B08, B11, B12 + quality bands
Sentinel-2 composite imagery 2021	EO (.tif)	B02, B03, B04, B05, B08, B11, B12 + quality bands
Tree Cover Density 2018	Auxiliary (.tif)	Tree canopy coverage, in %
Forest Type 2018	Auxiliary (.tif)	Forest type: Broadleaf, Conifer and non-forest
Forest Change 2000-2021	Auxiliary (.tif)	Year of detected tree cover clearance
strata_sets.csv	nFIESTA (.csv)	List of strata sets
strata.csv	nFIESTA (.csv)	Strata sets
cluster_configurations.csv	nFIESTA (.csv)	List of cluster configurations

panels.csv	nFIESTA (.csv)	List of panels
clusters.csv	nFIESTA (.csv)	List of plot clusters
plots.csv	nFIESTA (.csv)	List of plots
inventory_campaigns.csv	nFIESTA (.csv)	List of inventory campaigns
reference_year_sets.csv	nFIESTA (.csv)	List of reference-year sets
reference_year_sets_ panels_associations.csv	nFIESTA (.csv)	Possible combinations of panels and reference-year sets
plot_measurement_ dates.csv	nFIESTA (.csv)	List of plots with dates of their assessment
plot_cell_associations.csv	nFIESTA (.csv)	Spatial linkage of plots to the estimation cells
plot_target_data.csv	nFIESTA (.csv)	Plot-values of target variables
plot_target_data_ availability.csv	nFIESTA (.csv)	Missing target variables or categories
plot_auxiliary_data.csv	nFIESTA (.csv)	Plot-level values of auxiliary variables

5. Mapping and estimation module set-up

5.1 Module overview

The *Mapping and estimation* module is the core of the PathFinder platform where the production of high-resolution maps as well as statistically sound estimates is implemented to derive consistent European wide forest information down to 10 m spatial resolution on an annual basis. The module closely interlinks two components:

1. Map production with a combination of EO and field data.
2. Statistically sound estimation of target parameters in predefined geographical areas (cells), eventually complementing NFI data by auxiliaries in the form of wall-to-wall maps.

The close linkage and combined use of the maps and estimates is one of the biggest strengths and a unique feature of the PathFinder platform concept. This allows provision of consistent time series of forest estimates while also providing the latest forest information maps. In addition to serving as input for European forest scenario modelling conducted in WP 3 of the PathFinder project, the maps can also be used as inputs for improved statistical estimation (model-assisted or model-based). The vast field database used in the estimation also provides information for map uncertainty evaluation.

While the map production is conducted in Forestry TEP, the statistical estimation is run with nFIESTA. Both of these components have been described in detail in Section 3. For the functionality of the PathFinder platform, it is essential to ensure the connection between the two components, as well as the smooth input of datasets from the *Data preparation* module and output to the *Results* module. As outputs, the *Mapping and estimation* module provides 1) annual European wide forest attribute maps in 10 m spatial resolution and 2) annual forest statistics from arbitrary areas to national and European level.

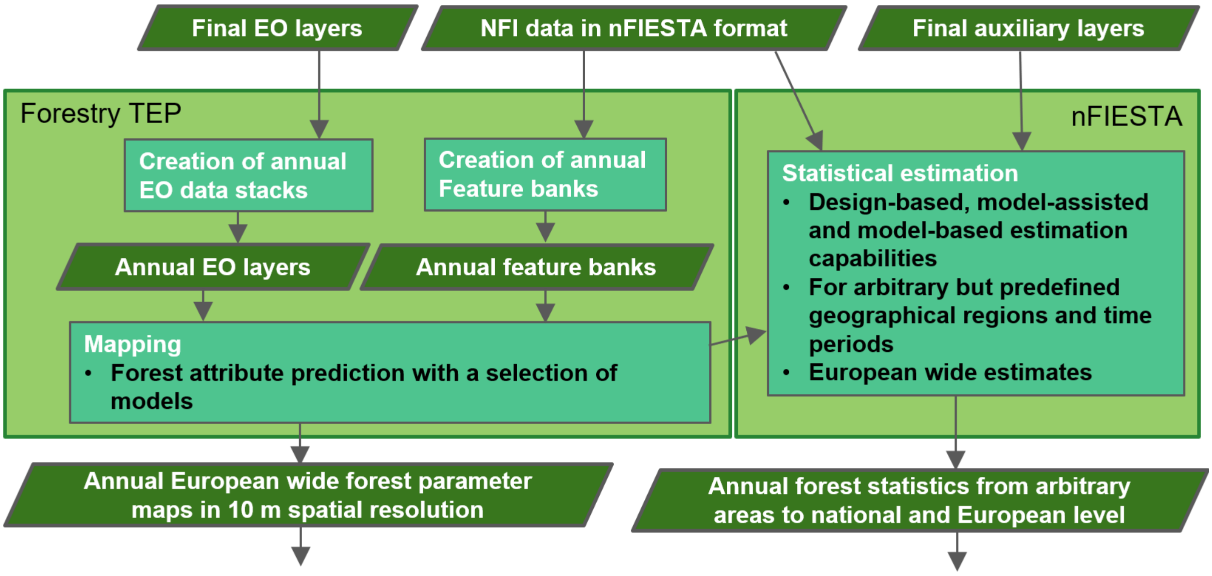

Figure 7. Overall workflow of the mapping and estimation module of the PathFinder mapping and estimation platform.

5.2 Mapping component

5.2.1 General overview and key characteristics

The main objective of the mapping component is to produce EO based timely maps using the field plot database with the most recent EO data. The mapping is based on a database of annual feature banks, which provides reference data for the mapping. The annual feature banks are created by combining harmonized field plot measurements (as described in the *Data preparation* module) with the EO data features acquired in the year of the measurement. The feature banks can be combined and used for any given year, assuming the consistency of the EO data feature values. In an operational setting, the collection of feature banks would be supplemented every year with measurements from the previous year. Whenever a new EO dataset is created, the latest available feature banks will be used in the map production. Note that in an operational setting, the quality of the feature bank database will improve rapidly during the first years, when more and more plot measurements will be linked to temporally coinciding EO feature values. The feature banks including the selected EO features will be used with a stack of EO layers including the same features. In the demonstration phase, only Sentinel-2 features are used (see Section 4) but the collection of EO features can be expanded in the future.

A wide range of methods can be implemented into the platform for map production. In the demonstration phase, the k Nearest Neighbour (k-NN; Alt 2001) method is used as the benchmark method. This method has been widely used for forest monitoring since the 1990's (Tomppo 1991; Tokola et al. 1996; Tomppo 2008; Chirici et al. 2016) and is operationally used for national level mapping in some European countries (Mäkisara et al. 2022; Reese et al. 2005). The flexibility of the platform allows utilization of other methods for map production as well. However, as the inter-annual and spatial consistency of the maps is of high importance, maps produced with different methods should not be used to draw conclusions regarding changes in forest attributes. During the platform demonstration, other methods may be tested and compared with the benchmark k-NN method (see section 5.2.3 for more discussion). One important issue that should be considered in the testing is the functionality of the mapping method in different types of forest ecosystems around Europe, ranging from large homogenous closed canopy forests to open canopy forest ecosystems with high species and structural variability.

5.2.2 Feature bank creation

One of the novel features of the PathFinder platform will be the growing feature bank database, combining harmonized plot level forest attribute values with contemporary EO data values. These feature banks are used in the map production. Initial plan for the feature bank creation workflow includes:

1. Calculation of national feature banks including: Plot identifier, year, Sentinel-2 tile, forest attribute values and EO feature values.
2. Error screening and outlier removal.
3. Splitting of feature banks to training and testing data sets.
4. Combination of national feature banks into European wide annual feature banks.

For the platform demonstration, a tool is implemented in Forestry TEP for feature bank creation. The tool is linked to nFIESTA. It reads in necessary csv files from nFIESTA and outputs annual csv files that include one row for each plot with desired forest attribute values and corresponding EO data values. The information required for the feature bank creation is included in the 'Plot_auxiliary_data.csv', 'Plot_cell_associations.csv' and 'Plot_target_data.csv' files. At this point the feature banks are created on national level.

The national feature banks are then screened for clear errors (caused e.g. by temporal mismatch between field measurements and EO imagery) and outliers are removed. The screened feature banks are split into training and testing feature banks, by extracting 1/3 of the plots for the testing feature bank that is set aside from the map production. The national feature banks are subsequently combined into annual European wide feature banks (both training and testing), which will form the basis for the growing repository of feature banks.

5.2.3 Map production

In the map production component, annual EO dataset and feature banks are used to produce forest attribute maps with associated error layers. These maps aim to:

1. Provide timely information on the latest development of forest resources in Europe down to 10 m spatial resolution.
2. Provide auxiliary information for model-assisted and model-based estimation of forest resources improving accuracy and/or availability of estimates in smaller geographical areas and in specific periods of time.

The first objective allows up-to-date monitoring of forest resources with a consistent method throughout Europe. The initial design of the system aims for annual maps to be published within the mapping year (e.g. 2020 growing season maps by the end of 2020). These maps would be based on the latest growing season EO data and the cumulative feature bank (not including the latest growing season plot measurements, however). Technically, maps could also be produced for any *ad hoc* purposes (e.g. evaluation of the effects of major natural disasters) with the most recent EO data. The second objective improves the accuracy of the model-assisted estimation to be conducted for small areas with insufficient number of field plots.

In the platform demonstration, the European wide benchmark maps are produced with the k Nearest Neighbour (k-NN) method (Alt 2001). The first version of a Forestry TEP tool has already been created for the implementation of the mapping. The tool requires as input the relevant EO data layers as well as the training and testing feature banks. In the demonstration the tool is using Sentinel-2 tiles, but it can technically handle different size images as well. In addition to the selection of output attributes, the tool allows the selection of key parameters of the k-NN approach, including most importantly the distance metric, number of neighbours and weights used in the prediction. In addition to the forest attribute prediction layers, standard deviation layers are produced based on the variance of the nearest neighbours used in the prediction. In addition to the graphical user interface (Figure 8), the tool can be used through a REST API connection which enables 1) easy execution of mass processing of hundreds of Sentinel-2 tiles and 2) linking of the tool to other platforms.

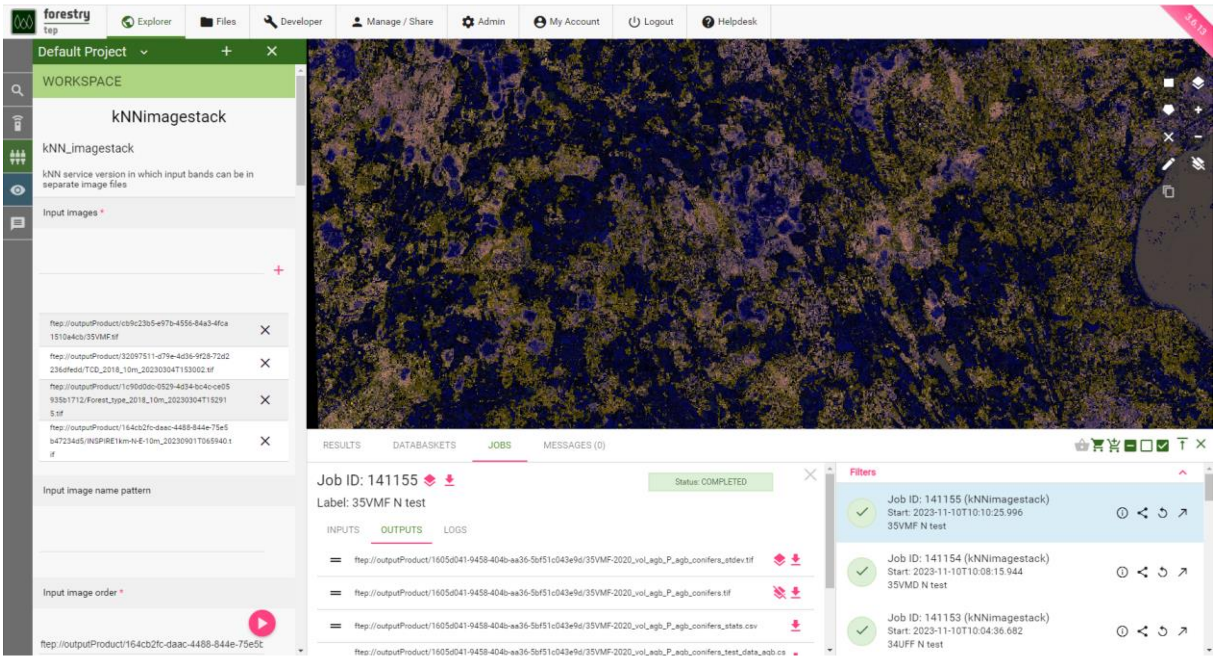

Figure 8. k Nearest Neighbour tool in Forestry TEP, to be used in the demonstration for the map production.

The tool will be further developed before the demonstration mapping with regards to the selection of the plots used in the prediction based on their geographic location. The approximate location provided in the feature banks will be used to select the optimal set of plots to be used in a given area, taking into account the varying density of plots in different parts of Europe (including countries with no plots available). In an operational context, similar selection strategies need to be defined also in the temporal space. The number of plots in the feature bank will increase annually, allowing the selection of optimal number of most recent plots, which may again vary in different parts of Europe based on field sampling schedules. The platform demonstration will provide valuable input for optimal plot selection, taking into account both the geographic and temporal aspects.

After the k-NN tool, the raw output maps subsequently pass through postprocessing tools which output single layer forest attribute maps in 10 m spatial resolution in the desired projection and tile size. Each forest attribute layer map is accompanied with a corresponding error layer, providing the standard deviation of the predictions (based on the variance of the seven neighbours used in the demonstration map production). The standard deviation layer will be used to accompany the predictions with an uncertainty interval.

Although the k-NN method is used in the demonstration, it does not mean that the system would be limited to using this method in the future. In principle, any other model linking the field data with the remote sensing data (such as parametric and machine learning models) can be implemented; the specific requirements of the platform for their implementation need careful consideration though.

5.3 Statistical estimation component

5.3.1 General overview and key characteristics

The production of statistical estimates of target parameters (totals or ratios, e.g. total forest area, or mean growing stock per hectare of forest) will be done by nFIESTA, namely its *nfiesta_pg* PostgreSQL extension (nFIESTA / nfiesta_pg · GitLab) controlled by the GUI module 'Estimates' (Home · Wiki · nFIESTA / nfiesta_gui · GitLab).

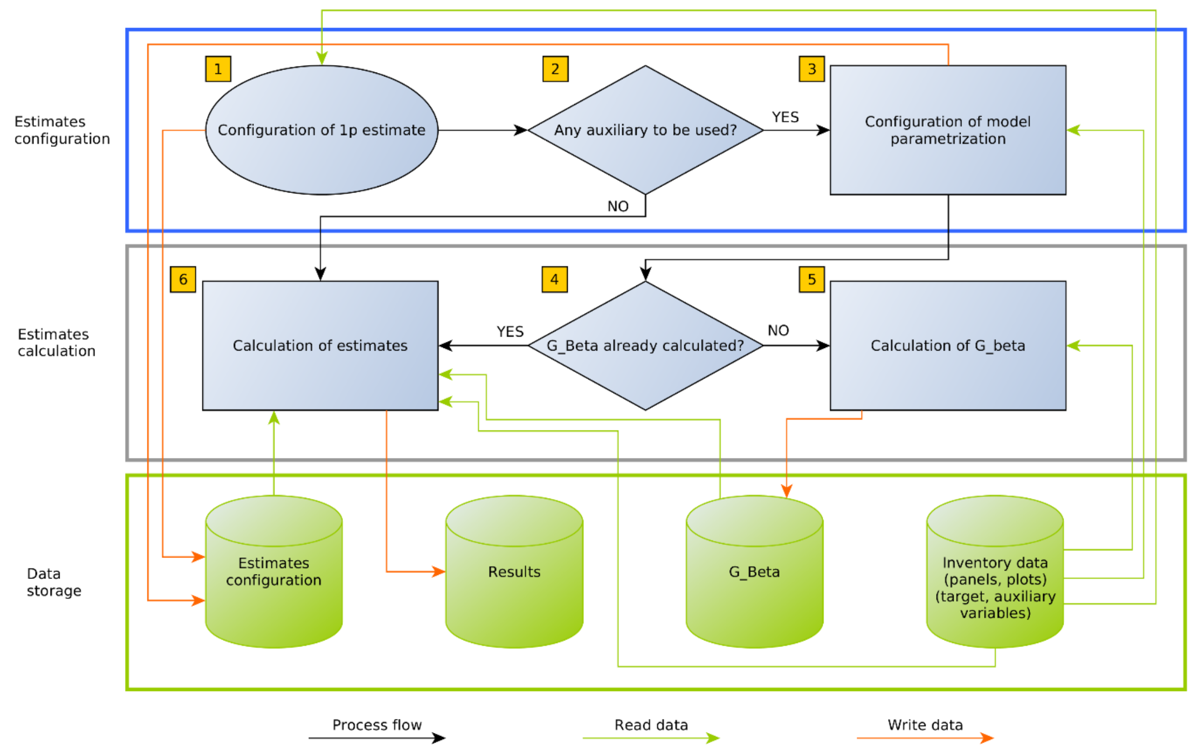

Figure 9: Functional schema of nfiesta_pg.

Figure 9 shows the three main functional parts of *nfiesta_pg*:

1. **Data storage** – which (i) contains pre-processed NFI and auxiliary data (uploaded either from CSV or directly from the *nfiesta_target_data* and *nfiesta_gisdata* extensions), (ii) saves configurations of various types of estimates by target parameter, estimation cell (geographical area) and time period, (iii) saves results of the estimations (point estimates of totals or ratios and their associated variances, minimal sample sizes and confidence intervals) and finally also saves parts (G_Betas) of parameterisations of working models (model-assisted or model-based estimations) that depend only on the configuration of auxiliaries, so they can be reused for estimation of more (new) target parameters (reduction of computation overhead).
2. **Estimates configuration** – for the estimation of a specific target parameter, in a specific cell and time period the optimal estimation technique (type of estimator) is identified according to the availability of NFI and auxiliary data. All the information needed for practical estimation is saved in the configuration – input data, model specification and parametrisation region.
3. **Estimates calculation** – which performs the calculation of selected estimators according to the saved configurations.

The following section contains detailed description of the configuration of estimates. More information about the other two functional parts of *nfiesta_pg* can be found in the report of Kohn et al. (2018).

5.3.2 Configuration of estimates

The key logic of *nfiesta_pg* is included in the configuration of estimates described by the schema in Figure 10 below.

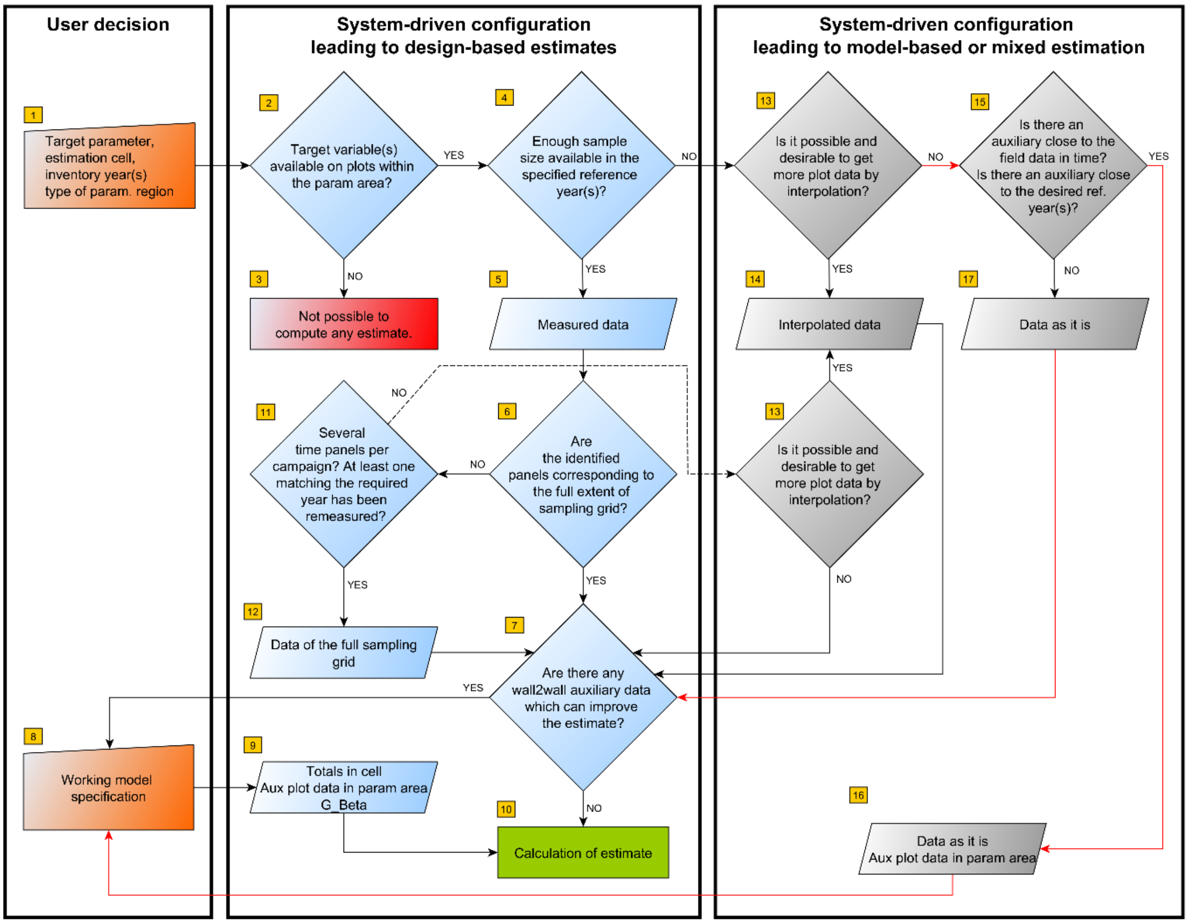

Figure 10. Configuration of estimation by nfiesta_pg

The configuration of the estimates formalises the search for the best possible data (field as well as auxiliaries if available) and the corresponding estimation technique that can be applied given the user specification of:

- **target parameter** – variable, e.g., total forest area.
- **estimation cell** – representing a geographical area for which the best possible estimator of the specified target parameter needs to be identified as well as the dataset allowing for its calculation.
- **period** – of time to which the estimate should (ideally) correspond.
- **parameterisation region** – a geographical region corresponding to the union of several cells, in which a working model can potentially be parameterised and used to increase the precision of the target parameter estimate.

These user inputs are marked by orange nodes in the very left part of the schema. The central part corresponds to design-based estimation, including model-assisted estimation that improves the accuracy of estimates by incorporating auxiliaries, but without relying on the validity of a model. These robust estimation techniques are always preferred by nFIESTA. Nevertheless, they can only be applied if the sample size for the desired combination of geographical area and time window is sufficient, meaning enough and up-to date NFI data must be available.

The right part of the schema leads to model-dependent or mixed estimators that come into play as last resort in situations when there is no timely matching field data in the estimation cell, or if the sample size is not large enough. But still, there must be enough field and auxiliary data in the parametrisation area to be able to fit and validate a model. In addition, the red arrows mark the purely model-based estimation scenarios that should be avoided whenever it is possible. The reason is that the quality of model-based estimates depends on the quality and validity of the model in use.

Proving model validity is typically demanding and not really feasible in a large production setting being considered by nFIESTA. Because models are specific for target variable and auxiliary combinations, they need to be developed and validated on case-by-case bases. It holds not only from the viewpoint of an analyst in charge of producing the estimates, but also from the end-user perspective. End-users typically do not have the capability, resources, time, or access to data needed for model validation. Therefore, if they are provided with model-based estimate only, they end up in a position either to trust or not to trust the analyst's choice of the model. They have no firm clues to tell how much the estimate can deviate from truth if the model was wrong in a specific way. The situation becomes even more difficult, if estimates produced by competing models considerably deviate, so alternative interpretations can be suggested.

The blue and grey diamonds mark decisions made by nFIESTA (potentially user-assisted decisions). The thin blue or grey rhomboids represent the dataset available for the calculation of estimates. Only if the green rectangle in the bottom part of the schema is reached, the configuration succeeded, and the estimate can be calculated. Reaching the red end point will stop the whole process because no configuration (estimation method) can be suggested for the user inputs.

To understand what estimation technique is suggested for what situation regarding NFI and auxiliary data availability the nodes of the schema in Figure 10 are described in detail.

- **Node No. 1** – the user chooses a target parameter (total or ratio and the target variable), a period, estimation cell (collection) and a parameterisation area by identifying a cell collection higher in the hierarchy compared to the chosen cell. The parameterization area is needed to test whether auxiliary data can be used for estimation or not.
- **Node No. 2** – automatic check whether the necessary target variable is available on all plots within the automatically identified parameterisation region. The check is done irrespective of the reference year.
- **Node No. 3** – because the requested target variable is not available in the parameterisation region, not even in parts of the region, no estimates can be produced. From here the user can come back to Node No. 1 and choose another cell collection, eventually leading to bigger parameterisation regions and try to pass through Node No. 2 again.
- **Node No. 4** – data identified in Node No. 4 is reduced to a subset corresponding to the desired reference year(s) within the estimation cell. The size of the subset is compared to the minimum sample size requirement stored in the *nfiesta_pg* setting.
- **Node No. 5** – plot data that matches spatially (panels) as well as temporally (reference year sets) to the input user specification in the Node No. 1.
- **Node No. 6** – system check whether the plot data compiled the Node No. 5 represents the full set of plots of an inventory campaign. The purpose of this check is to distinguish between

situations when only single-phase estimator can be used (NFI data matching the time period) and when preceding NFI data can be incorporated as auxiliary in the two-phase setting.

- **Node No. 7** – user assisted decision whether auxiliary data obtained from wall-to-wall maps can be used to improve estimation precision. A mapping between working models, auxiliaries and target parameters can be saved in the *nfiesta_pg* configuration, so the decision can be made automatically.
- **Node No. 8** – the user must specify auxiliaries and all the terms of a linear regression model which will be used for the estimation. The mapping mentioned in Node No. 7 can automate this step as well, by saving suitability rank for working models and auxiliaries that can be used with a target parameter.
- **Node No. 9** – the dataset necessary for the estimator type, i.e., G-beta matrices (Adolt, Lanz, & Fejfar, 2018, pp. 17, 21,22), auxiliary totals for the estimation cell, and auxiliary values for plots in the parametrisation area. The G-beta matrices can either be reused (if available) or they must be evaluated by nFIESTA at this moment.
- **Node No. 10** – nFIESTA calculates the respective estimate the type of which depends on the path through which this node has been reached.
- **Node No. 11** – system checks whether there are more panels representing various reference years in the dataset prepared 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 remeasured (measured also in the past). If both conditions are met a two-phase estimator can be constructed combining the data of panels for the requested period (taking the role of second-phase data) and data measured in the past (considered first-phase data), so all panels of an inventory campaign are included (either in the first or second phase data, some panels are included in both phases).
- **Node No. 12** – the first-phase and the second-phase data including all panels according to the NFI design (the full sampling grid) that will be used for two-phase estimation. Wall-to-wall auxiliaries may be added later, depending on the answer in Node No. 7.
- **Node No. 13** – either there was no suitable data found in Node No. 4 or the data does not correspond to the full sampling grid (all panels) and, at the same time, two-phase estimation (Node No. 11) is not possible. In the first case one can try an interpolation of plot values between two measurement years or using the unchanged plot values corresponding to the preceding survey (if an interpolation between two field survey occasions is not possible). This corresponds to model-based or mixed inference. In the second case (reaching Node No. 13 from Node No. 11), the default is to continue to Node No. 7 (avoiding interpolation because the amount of NFI data may be sufficient, a design-based estimation is preferred).
- **Node No. 14** – the system creates a plot dataset which contains the original as well as some interpolated panel (between two measured reference year sets). 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 estimates. Estimates using interpolated data should not be considered design-unbiased.
- **Node No. 15** – a situation of no NFI data measured in specified reference year(s) and no interpolation or updating possible. Check if there are wall-to-wall auxiliary data corresponding to the measurement year(s) of field data of 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 NFI data followed by a synthetic (model-based) estimation using the fitted model with auxiliaries corresponding to the user requested reference year(s).

- **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). This situation could eventually be circumvented by enlarging the parameterisation area.

Positive answers are given only in case the conditions are met for each sampling stratum intersecting the parameterisation region. This basically means that we configure the same type of estimate in all strata found within the given estimation cell. The advantage is ease of interpretation of accuracy estimates – an estimate for a cell is difficult to understand if the way it was obtained mixes design-based and model-based inferential concepts.

5.3.3 Technical implementation for the platform demonstration

Current version of *nfiesta_pg* implements estimators corresponding to paths 1-2-4-5-6-7-10 (design-based, single-phase estimation with no auxiliaries) and 1-2-4-5-6-7-8-9-10 (single-phase estimation with wall-to-wall auxiliaries). The second path may correspond either to:

- **Model-assisted (design-based) estimation** – if the actual sample size within the estimation cell is not zero. So, the statistical inference (unbiasedness and variance) does not require model validity.
- **Model-based (synthetic estimator)** – if the actual sample size in the cell is zero (assuming the corresponding limit in Node No. 4 is set to zero too), but a model can still be parameterised using data in the parameterisation region. Statistical inference assumes model validity. For a specific group of working models, the estimator is also model-assisted at the level of the whole parameterisation area. Therefore, the closer the parametrisation areas are to estimation cells, the more robust is this class of estimators.

The estimators corresponding to these two paths are described in the technical report of Adolt et al. (2018). However, the nFIESTA GUI module only implements single-phase estimation with no auxiliaries. Single-phase estimation with wall-to-wall auxiliaries will be added to the GUI and demonstrated within the Pathfinder project. Estimators corresponding to the remaining schema paths are specified in the report of Ene et al. (2018).

5.4 Mapping and estimation module input and output datasets

The inputs for the *Mapping and estimation* module are the outputs from the *Data preparation* module described in Table 4. To retain the “stand-alone” nature of the different modules in the PathFinder platform concept, it is essential to ensure that the standard data formats following the OGC protocols are used. For example, in the case of the platform demonstration, EO datasets could have been produced in any other platform and only downloaded to the Forestry TEP for processing. This allows flexibility of the operational implementation of the platform, taking advantage of available infrastructures.

The outputs of the *Mapping and estimation* module in the platform demonstration include maps and statistics (Table 5). The forest attribute maps to be produced depend on the available datasets and used mapping methods. In the PathFinder platform demonstration, at least the following attribute maps will be produced: Stem volume, above ground biomass, conifer proportion (of biomass). All of the attribute maps will be accompanied with a corresponding error layer, providing the standard deviation of the prediction. The maps will have European wide coverage and they will be provided in the European Terrestrial Reference System 1989 (ETRS89) in Lambert Azimuthal Equal Area (LAEA) projection in 10 m spatial resolution.

Table 5. Mapping and estimation module output datasets.

Dataset	Type	Characteristics
Forest attribute maps	map (.tif)	Annual European wide maps of forest attributes in 10 m spatial resolution
Forest attribute error maps	map (.tif)	Standard deviation of the forest attribute predictions in 10 m spatial resolution
Estimation component outputs	DB (all), CSV (estimates only), JSON (metadata only)	From nFIESTA GUI estimates can currently be exported to CSV file with a custom selection of included columns. However, the CSV does not contain metadata of local densities (NFI plot-level data). This is highly structured and saved in JSON in the DB. The intention is that the estimates are made accessible through *nfiesta_results* extension to which other services can connect. If needed, the export of JSON can be bound to the export of CSV from nFIESTA GUI.

6. Results module set-up

6.1 Module overview

The main objective of the *Results* module is to provide all stakeholders access to and easy visualization of the PathFinder platform output products. A dedicated PathFinder visualization and distribution portal will be the main entry point for stakeholders interested in using the platform. This portal will allow visualization and download of all the output products and publicly available input datasets.

In addition, the aim is to create automated linkages to other platforms interested in a regular feed of outputs from the PathFinder portal. These platforms could include e.g. the Forest Information System for Europe[16] (FISE) or national databases of countries interested in using the datasets produced in the PathFinder portal for their own analyses or reporting.

Initial plans of the PathFinder portal are provided below, followed by discussion on potential linkages to FISE or national platforms. During the platform demonstration in the project phase, only the PathFinder visualization and distribution portal will be technically implemented and demonstrated.

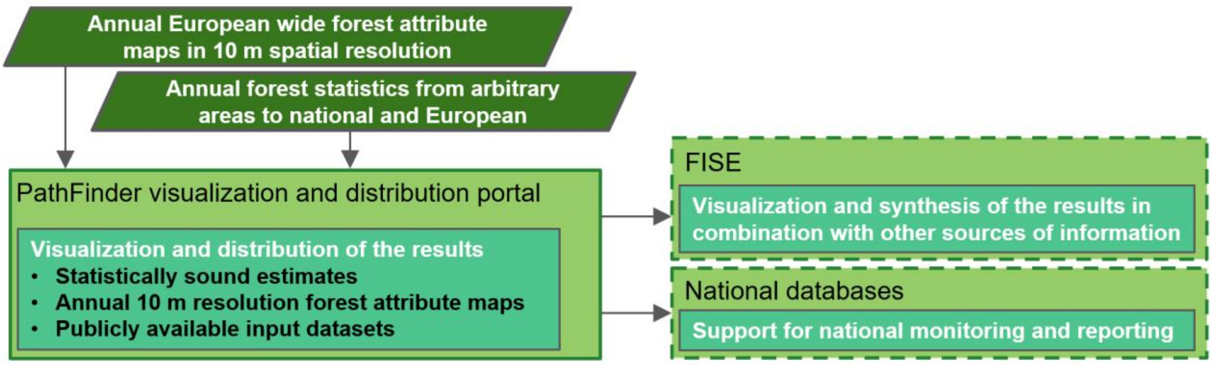

Figure 11. Overall workflow of the results module of the PathFinder mapping and estimation platform.

6.2 PathFinder visualization and distribution portal

The main entry point to access the PathFinder platform datasets will be the PathFinder visualization and distribution portal. Initial plans of the portal design include the following key concepts:

[16] https://forest.eea.europa.eu/

- Geographic Information System (GIS) based approach.
- Search functionalities to derive estimates and maps for geographic areas of interest.
- Visualization of historical development of estimates and maps in geographic areas of interest.
- Tabulation of estimates for administrative, INSPIRE grid (European Commission, 2024) or arbitrary defined interest areas and time periods.
- Download of maps, publicly available input datasets and forest statistics tables.

The portal will have a GIS based approach that enables visualization of the input EO and output map layers as the basis. The portal will be set up on a publicly availably server at the Cloudferro[17] infrastructure, where Forestry TEP is also located. The search functionalities and estimate calculations will be based on the European Nomenclature of territorial units for statistics[18] (NUTS) regions and an INSPIRE grid structure.

The user will be able to visualize the input EO data, output maps (Figure 12) and the forest attribute estimates for the area in question (Figure 13). The forest attribute statistics (e.g. above ground biomass) will be available in tabular form as well as in graphical form. Furthermore, it should be possible to visualize the historical development of the forest attribute statistics in a given area.

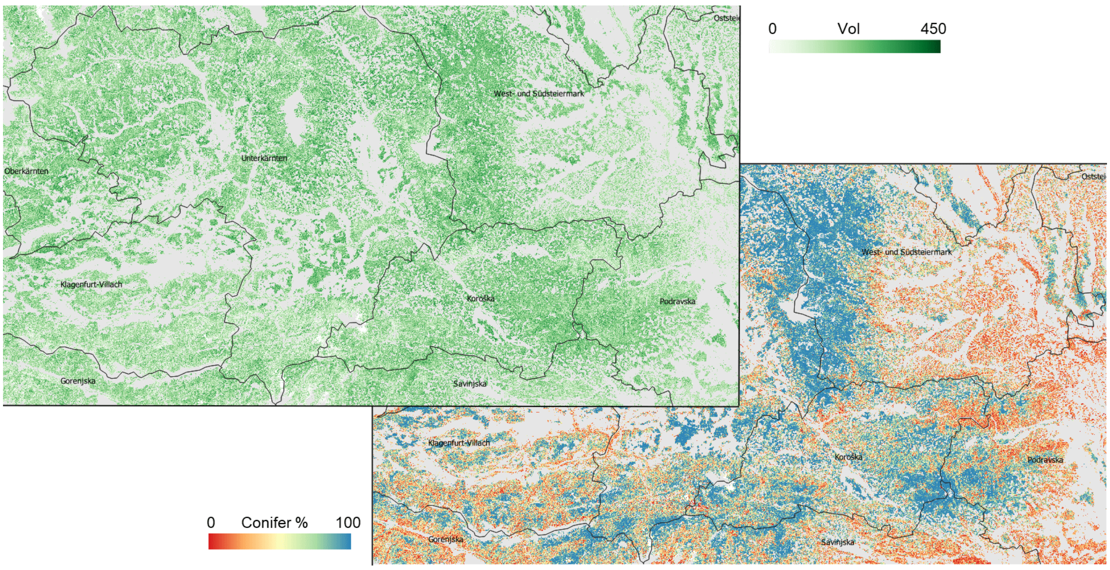

Figure 12. Example of output map layers to be provided in the PathFinder platform. Volume (m^3/ha) and conifer proportion (% of biomass) in Central Europe.

[17] https://cloudferro.com/en/
[18] https://ec.europa.eu/eurostat/web/nuts/background

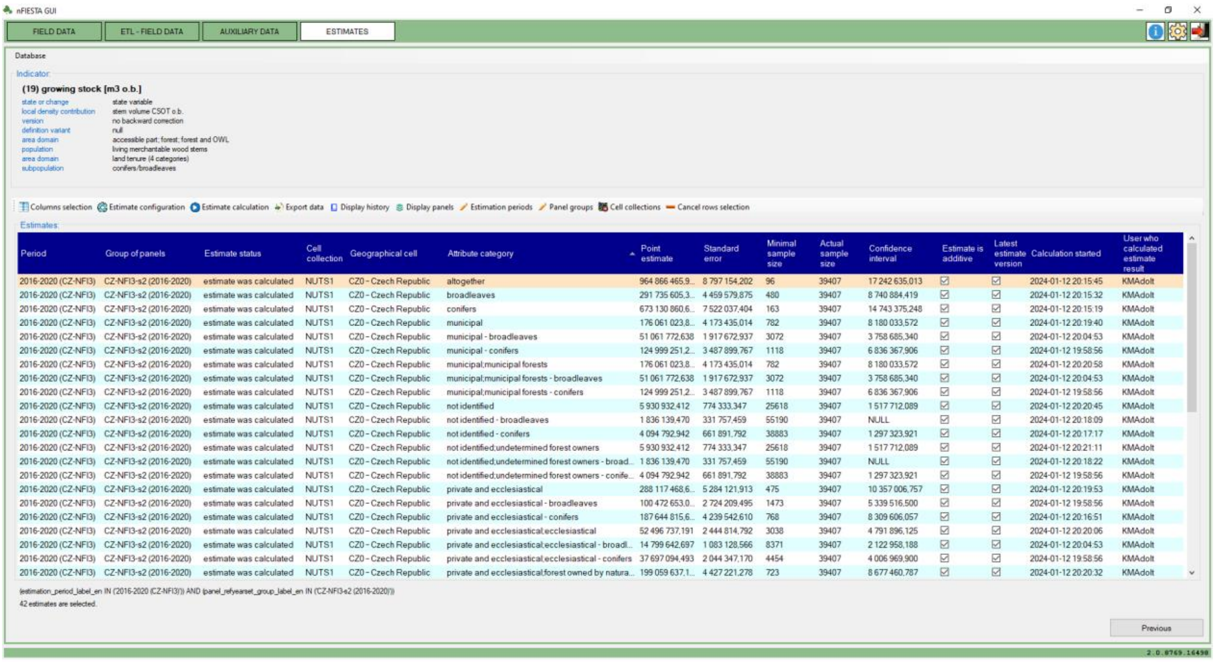

Figure 13. A screenshot showing outputs of the nFIESTA estimation module.

The user should be able to download all publicly available underlying input data (e.g. EO and auxiliary layers), as well as the output maps and estimates. The forest attribute estimates will be provided in tabular format for predefined areas of interest.

6.3 Forest Information System for Europe (FISE) and national databases

As the PathFinder project aims to develop and demonstrate an innovative integrated forest monitoring and pathway assessment system that allows a consistent EU wide analyses, it is essential that the system is integrated to existing and future infrastructures and practices. The forest information produced in the PathFinder platform provides valuable inputs for forest policy definition and scenario forecasting. It is therefore essential that the information is available in timely manner for all stakeholders potentially benefitting from it. Potential stakeholders interested in automated feeds from the system include for example:

1. Forest Information System for Europe (FISE)
2. National databases

FISE is a platform developed by the European Environment Agency[19] (EEA) for sharing information with the forest community on Europe's forest environment, its state and development. FISE brings together data, information and knowledge gathered or derived through key forest-related policy drivers. FISE is dependent on data and information coming from the EU and EEA member states. The role of FISE is expected to grow in the implementation of the harmonized European wide forest monitoring framework outlined in the New Forest Strategy 2030.

In this context, it would be important to enable an automated feed of selected datasets from the PathFinder platform (when operational in the future) into FISE. Discussion on what datasets would be useful for FISE need to be started during the platform demonstration to optimize the technical characteristics of the selected datasets to best meet the needs of FISE. In longer perspective, issues concerning for example quality control, operational product production cycles and management of

[19] https://www.eea.europa.eu/en

operations would also need to be settled. However, these issues will go beyond the project as they are strongly connected to the operational implementation of the PathFinder platform.

As the requirements on member countries related to monitoring and reporting of forest and environmental EU regulations are growing rapidly (EC 2023), the harmonized products from the PathFinder platform may become highly useful for national level organizations for monitoring and reporting purposes. Approaches very similar to the ones used in the PathFinder platform are operationally used in some EU member countries. The countries that currently do not use such approaches or do not have access to these types of products, may greatly benefit from the use of the PathFinder products.

If interest in the use of the PathFinder products is expressed by the member states, the platform should be able to facilitate the use of the products with automated connection to national databases for easy and timely access to the output products. Creation of products tuned for a given country could also be considered as the system technically allows creation of products with national datasets and definitions as well. Overall, it will be important to ensure that the capabilities and outputs of the PathFinder platform will be fully used by interested member countries and the access to the system will not become a limiting factor. Potential interest and requirements by member states as well as the details of the potential direct links to national databases will need to be discussed in the context of the platform demonstration in preparation for the operational system set-up.

7. Future development directions

A basis for a future forest monitoring platform is designed and demonstrated during the PathFinder project. The main goal of the demonstration is to prove the scientific and technical feasibility of the system. The flexible implementation approach leaves room for further development within the platform modules and for the operational implementation of the system.

In this section we discuss key areas of further development possibilities, beyond the scope of the PathFinder project. Some of the points raised below relate to management and operating decisions that need to be made, while other points relate to topics that are under constant rapid development (e.g. EO sensors).

1. *Operational set-up*

It is important to emphasize that this document has concentrated only on the scientific and technical aspects of the PathFinder platform design, outlining the plans for platform demonstration to be conducted within the PathFinder project. Beyond the scientific and technical feasibility, implementation of the system for operational use requires major infrastructure and management decisions to be made. The operational set-up may vary significantly depending on the entities selected to manage the platform.

There are numerous Information and Communication Technology (ICT) infrastructures in Europe that can technically host the entire system or parts of it. The flexible module design allows implementation of the components into different infrastructure (as is the case during the PathFinder project). A key barrier for setting up any centralized storage and processing system for the proposed platform concept is the storage of sensitive field plot locations. If agreement on a single infrastructure to host all of the input data (including the sensitive field plot locations) cannot be reached, field data pre-processing and EO feature extraction will need to continue to be handled by the country correspondents in the same manner as it is done in the PathFinder platform demonstration. This may to some extent limit the potential of further development. See point 3 below for more discussion on this topic.

Management of the operational system is perhaps the most important issue affecting the operational set-up. The entities selected to manage the system will ultimately decide on the details of the scientific and technical set-up of the operational implementation based on the management rules and regulations defined for the system. The upcoming European forest monitoring law which is currently under discussion (EC 2023), may also have an impact on the management of the operational system, if e.g. stakeholders come together to set up a management structure for a system to support countries to provide the reporting required in the upcoming law. Therefore, no decisions on the implementation of the operational system have been made or can be made at this point, which further underlines the importance of the flexibility of the technical solutions chosen for the demonstration conducted in the PathFinder project.

2. EO and field data feature banks

The feature banks (i.e. the sets of EO values and corresponding field plot target variables) created by the system would be valuable also beyond their use in the PathFinder platform. Hundreds of thousands of field measurements combined with corresponding EO and auxiliary values collected on annual basis will form an unprecedented reference database for EO based forest monitoring studies in Europe. However, in order to be able to make the feature banks public, it would be essential to preserve the plot location privacy.

The current approach used in the feature bank creation theoretically enables reverse engineering of the plot location. If, through some (EO) data modifications, it could be guaranteed that this is not possible, a separate public version of the feature banks could potentially be created in the operational phase, leaving out the approximate plot location. However, the necessary modifications may also significantly reduce the usefulness of the feature banks. Possibilities to create feature banks that would ensure plot location privacy but still maintain the usefulness of the data will be investigated during the platform demonstration phase.

Nevertheless, alternative solutions for integration of the field and EO datasets should continue to be envisioned. Depending on the mapping methods used, access to the actual field dataset by the people producing the maps may not always be needed. Models could be created in a secured server, outputting only the necessary parameters for map production, without the need to access the database used for the model creation. However, this approach would need establishment of a centralised secured server which would host all the EO and field data used in the model creation. Management and hosting of such a server would need to be agreed upon. Accessing the field plots directly in national databases in this manner would not remove the hindrance of EO data download into all national databases. Furthermore, this approach would not work with all mapping approaches, as many methods (like the k-NN used during the project) require access to the feature bank in the mapping phase.

Overall, establishment of an effective and secured solution for accessing and utilizing the field data without jeopardizing the plot location in the future would help to fully unleash the potential of the vast NFI field plot databases and expanding EO datasets also beyond the PathFinder project.

3. Expansion of available EO and auxiliary datasets

Only Sentinel-2 optical satellite imagery is used in the PathFinder project demonstrations. Although it is a very powerful dataset for forest monitoring, enabling expansion of EO data sources in the operational system is essential for maintaining the relevance of the system in the future.

Firstly, inclusion of different types of space borne EO datasets may improve the mapping results. For example, integration of L-band Synthetic-Aperture Radar (SAR) datasets and spaceborne Light Detection and Ranging (lidar) data into the system should be investigated. Several new L-band SAR satellites (e.g. NISAR, ALOS-4 and ROSE-L) are planned in the near future and may help to improve the mapping accuracy with their canopy penetrating capability. Initial investigation of L-band radar is planned already during the PathFinder project with the new PathFinder field plots. Spaceborne lidar data (e.g., from ICESat-2) may also be useful in the future, although currently not providing wall-to-wall coverage.

Apart from space borne datasets, integration of airborne laser scanning (ALS) data in the system will be important in the future. The availability of ALS data in Europe is increasing rapidly. In the future, the operational European forest monitoring system should be able to also utilize ALS data where available.

Furthermore, the availability of auxiliary datasets potentially useful for the estimation will increase in the coming years. For example, the Copernicus Land Monitoring Service portfolio[20] will expand significantly, including, e.g., yearly maps of the Tree Cover Density, Forest Type and other products. These and other layers may provide valuable auxiliary information to support the model-assisted estimation conducting in the PathFinder platform. However, their usability for the mapping will remain unlikely as the aim of the platform is to produce up-to-date maps. The auxiliary layers for the given year will not be ready by the time of the mapping.

Overall, in order to retain the relevance of the system into the future, it is important to keep updating the selection of EO and auxiliary datasets as new and improved datasets become available. Attention needs to be given to maintaining the consistency of the results over time while new data sources are integrated into the system. In case of a potential operational system in the future, decisions on the datasets to be used need to be regularly made according to the platform management rules and regulations.

4. Improved management of NFI data and estimations

A meaningful monitoring of European forests can be implemented if NFIs cover the whole continent with a sufficient temporal resolution. If the field data acquisition frequency was less than annual, synchronisation of actual data acquisition years among all NFIs would become an issue. It would be far from ideal if countries delivered NFI data (not results), e.g., once in five years, but at the same time, if the actual acquisition year varied among the countries.

It is not only the NFI data collection in the countries itself that poses a challenge for the European forest monitoring implementation. Working on the presumptions that

- NFI data might (finally) be needed on yearly basis,
- NFIs (countries) will keep their current autonomy in terms of the survey design,
- NFIs keep their own technological solutions,

[20] https://land.copernicus.eu/en/products

it becomes very clear, that a robust technical solution for NFI data integration and analysis is needed.

The NFI data preparation according to a common specification, transfer and quality checks is a lengthy, technically demanding process. Every time, it took several months until the NFI data of several countries could be centrally analysed in nFIESTA (experience from DIABOLO as well as PathFinder). However, the duration of this process should not exceed a few weeks for proper operationalisation of the European forest monitoring system. At this stage, it is not the nFIESTA that slows down this process, but the complexity and variability of NFI survey designs, national databases, and specific demands on harmonising the plot data and the personal resources in the NFI institutions. Note that variables must follow a set of common definitions, harmonised across EU countries. The description of each single NFI design must be attached to every NFI dataset managed by nFIESTA. Otherwise, no sound statistical estimates can be produced and delivered to their users.

The data preparation, transfer and checking process can proceed significantly faster, if nFIESTA can be attached closer to the national databases, so the plot level data can be produced in a standardised, well-documented way by each country (metadata describing the nature of the data as well as key steps of its processing). This would of course need training of the NFI data providers as well as consultancies how to establish connections between the national DBs and nFIESTA (*nfiesta_pg* extension, local connection). The GUI module 'Field Data' will certainly need additional development to increase user comfort and eventually also include new scenarios, in response to the growing experience with its use.

Alternatively, the countries themselves would develop their solutions to be able to prepare the required datasets in standard format, without errors and in time. Every country would spend resources for functionality, that is to some degree common for all. In addition, the development of the nFIESTA (if used centrally at the EU level), might not be so convenient and straightforward if the NFIs' self-developed solutions would have to accommodate the changes as well.

When it comes to estimation methods, the corresponding *nfiesta_pg* and GUI module 'Estimates' need more development, so all paths (major classes of estimation techniques) presented in the diagram in Figure 10 (section 5.3.2) are covered.

Current modules of nFIESTA should be considered working prototypes in the sense that they were once implemented and remained in use without major revisions. We are aware of slow computations in some parts of the nFIESTA functionality. These may become a serious bottleneck if the DBs are filled by data of many NFIs or if nFIESTA is intensively (concurrently) used by many users for several years, causing the database size to increase a lot. We have in mind several improvements of performance. These often go into the depth of DB design of PostgreSQL extensions and consequently impact nFIESTA GUI modules that may need at least partial reimplementation (due to a possible API change).

Unfortunately, with the current number of nFIESTA developers and their multiple duties (not limited to nFIESTA development only) such a revision has not been possible, and the emphasis has been on the implementation of all key parts of the NFI data processing pipeline. However, nFIESTA's performance reimplementation should be considered in case of serious plans of using it in a (big) production environment.

References

Adolt, R., Fejfar, J. and Lanz, A. 2018. nFIESTA (new Forest Inventory ESTimation and Analysis) - Estimation methods implemented by the nFIESTA software. In 'Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales'. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Alt H. 2001. The Nearest Neighbor. In: Alt H, editor. Computational Discrete Mathematics. Vol. 2122. Berlin, Heidelberg: Springer Berlin Heidelberg. (Goos G, Hartmanis J, Van Leeuwen J, editors. Lecture Notes in Computer Science). p. 13–24. [accessed 2023 Jun 28]. http://link.springer.com/10.1007/3-540-45506-X_2

Astola, H., Häme, T., Sirro, L., Molinier, M. and Kilpi, J. 2019. Comparison of Sentinel-2 and Landsat 8 imagery for forest variable prediction in boreal region. Remote Sensing of Environment 223: 257–273. doi: 10.1016/J.RSE.2019.01.019

Chirici, G., Mura, M., McInerney, D., Py, N., Tomppo, E.O., Waser, L.T., Travaglini, D., and McRoberts, R.E. 2016. A meta-analysis and review of the literature on the k-Nearest Neighbors technique for forestry applications that use remotely sensed data. Remote Sensing of Environment 176: 282-294. doi: 10.1016/j.rse.2016.02.001.

EC 2023. European Commission proposal for a new monitoring framework for resilient European forests, COM 2023 728. Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52023PC0728&qid=1707223205537

Ene, L.T., Lanz, A., Adolt, R., Fejfar, J., Kohn, I., Morneau, F., Riedel, T et al. 2018a. Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Ene, L.T., Lanz, A., Adolt, R. and Fejfar, J. 2018b. A selection of methods proposed for the forest information and estimation system. In 'Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales'. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Hansen, M.C., Potapov, P.V., Moore, R., Hancher, M., Turubanova, S.A., Tyukavina, A., Thau, D., Stehman, S., Goetz, S.J., and Loveland, T.R. 2013. High-resolution global maps of 21st-century forest cover change. science 342(6160): 850-853.

Kohn, I., Adolt, R., Fejfar, J. and Lanz, A. 2018. nFIESTA (new Forest Inventory ESTimation and Analysis) - Functional description of the software. In 'Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales'. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Lanz, A., Adolt, R., Fejfar, J., Kohn, I., Morneau, F., Pesty, B. and Riedel, T. 2018a. Demonstration study based on NFI plot data and large-area, high-resolution auxiliary data. In 'Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales'. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Lanz, A., Adolt, R., Kohn, I. and Fejfar, J. 2018b. Specification of the eFOREST 2.0 CSV data file formats. In 'Report on the development of imputations techniques and updating algorithms, including concept and implementation and demonstration on studies at various time points and spatial scales'. https://gitlab.com/nfiesta/nfiesta_pg/-/wikis/myuploads/d2.10_report/d2.10_report.pdf

Miettinen J., Carlier S., Häme L., Mäkelä A., Minunno F., Penttilä J., Pisl J., Rasinmäki J., Rauste Y., Seitsonen L., Tian X., Häme T. 2021. Demonstration of large area forest volume and primary production estimation approach based on Sentinel-2 imagery and process based ecosystem modelling. International Journal of Remote Sensing 42:24, 9492-9514, doi: 10.1080/01431161.2021.1998715

Mäkisara, K., Katila, M. & Peräsaari, J. 2022. The Multi-Source national forest inventory of Finland — methods and results 2017 and 2019. Natural resources and bioeconomy studies 90/2022. Natural Resources Institute Finland, Helsinki 73 p

Reese, H., Granqvist-Pahlén, T., Egberth, M., Nilsson, M., & Olsson, H. 2005. Automated estimation of forest parameters for Sweden using Landsat data and the kNN algorithm. Proceedings of the 31st International Symposium on Remote Sensing. June 20–24, 2005. St. Petersburg, Russia. http://www.isprs.org/proceedings/2005/ISRSE/html/papers/652.pdf.

Tokola, T.J., Pitkänen, S. and Muinonen E. 1996. Point Accuracy of a Nonparametric Method N Estimation of Forest Characteristics with Different Satellite Materials. International Journal of Remote Sensing 17 (12): 2333–2351.

Tomppo, E., Olsson, H., Ståhl, G., Nilsson, M., Hagner, O., Katila, M. 2008. Combining national forest inventory field plots and remote sensing data for forest databases. Remote Sensing of Environment 112, 1982–1999.

Tomppo E. 1991. Satellite image-based national forest inventory of Finland. International Archives of Photogrammetry and Remote Sensing 28: 419-424.