Metsätehon raportti 240
21.12.2016

Data platform promoting forest data utilization through uniform access to heterogeneous data

DIGILE's Data to Intelligence (D2I) program

Miika Rajala
Risto Ritala

ISSN 1796-2374 (Verkkojulkaisu)

METSÄTEHO OY
Vernissakatu 4
01300 Vantaa

www.metsateho.fi

Data platform promoting forest data utilization through uniform access to heterogeneous data

DIGILE's Data to Intelligence (D2I) program

Miika Rajala
Risto Ritala

Metsätehon raportti 240
21.12.2016

ISSN 1796-2374 (Verkkojulkaisu)

SISÄLLYS

TIIVISTELMÄ

Metsävaratieto on paikkatietoa, joka on hajallaan monissa järjestelmissä. Lähitulevaisuudessa metsävaratiedon tuottajien joukko monipuolistuu entisestään: metsäkoneet nähdään merkittävinä big datan lähteinä ja on uskottavaa, että kansalaiset, kansalaisjärjestöt, viranomaiset ja yritykset keräävät yhä enemmän metsistä havaintoja ja mittaustuloksia, jotka auttavat arvioimaan metsävaroja nyt ja ennustamaan niiden tulevaa kehitystä. Tämän tiedon hyödyntäminen on sitä tehokkaampaa mitä helpommin sovellukset voivat saada datalähteet käyttöönsä ja yhdistää datoja entistä tarkemmiksi estimaateiksi.

Tässä raportissa määritellään palvelualusta (”platform”), joka saa heterogeenisen metsävaratiedon näyttämään sovelluksen kannalta homogeeniselta tietokannalta ja joka tarjoaa datan yhdistämisen ja datan käyttöoikeuksien palveluja. Palvelualustan ajatellaan tukevan muun muassa puukauppaan, metsäsuunnitteluun ja työmaasuunnitteluun liittyviä sovelluksia, mutta ennen kaikkea se laskee kynnystä kehittää monimuotoisesti metsävaratietoa hyödyntäviä digitaalisia palveluita.

Palvelualustan teknisenä ytimenä on kyselyrajapinnan määrittely sekä tietorakenne metsävaratiedon ja siihen liittyvän metadatan esittämiseksi. Metadatan olennaisena osana on tieto mittausdatan epävarmuudesta. Epävarmuustiedon perusteella kohdesuuretta pystytään estimoimaan yhdistämällä monen tietolähteen havaintoja soveltaen bayesilaisen datafuusion menetelmiä: yhdistetyt estimaatit ovat aina yksittäisen lähteen tietoja tarkempia.

Koska kohdesuuretta koskevat havainnot eri datalähteissä ovat tyypillisesti eri ajanhetkiltä, palvelualusta tarjoaa mahdollisuuden liittää datan käsittelyyn esimerkiksi puuston kasvumalleja tiedon ajantasaistamiseksi. Palvelualusta on varsin yleinen paikkatietojärjestelmä, joten siihen voidaan myöhemmin liittää metsävaratiedon lisäksi metsän käytön kannalta olennaista olosuhdetietoa.

Raportissa esitetään yksinkertainen, tietorakennetta, ajantasaistusta ja datafuusiota testaava esimerkki, jossa käsitellään todellista metsävaradataa.

Tutkimus kuului DIGILE:n Data to Intelligence (D2I) -tutkimusohjelmaan (2012–2016).

ABSTRACT

This paper considers the definition and requirements of a platform for providing data inquiry services for users and applications to easily access available forest data sources. For the applications to have a uniform view to varying heterogeneous forest data sources we specify a common data inquiry interface and a data structure for representing data and required metadata, in particular, the uncertainty. Furthermore, to make the access and use of the data sources simple, we derive the basic principles of predicting data sources up-to-date with growth prediction models, and in particular, for combining several up-to-date data estimates by means of Bayesian data fusion. The platform functions as the basis for new applications, and is expected to enhance businesses productivity of the forest sector, e.g., related to wood trade, forestry planning and worksite planning. We will demonstrate the platform with a simple real forest data case for testing the data structure, and the data updating and data fusion services. The study was carried out as a part of DIGILE's Data to Intelligence (D2I) program (2012–2016).

Key words: platform, inquiry interface, data abstraction, data structure, Bayesian data fusion

1 INTRODUCTION

Forest resources today can be inventoried very accurately; airborne or terrestrial LiDAR (Kankare et al. 2014) based methods can be used even for constructing maps of individual trees and their properties (Raumonen et al. 2013). There are many other methods available, such as imaging or satellite radar based methods (Holopainen et al. 2010, Antropov et al. 2013, Vastaranta et al. 2014) Varying methods provide information on trees with varying time cycles, resolutions, accuracies, and costs.

Some of the methods are directly competing with each other, while others are rather completing each other. For example, cost-effective short-cycled satellite-based data has potential in updating costlier longer-cycled LiDAR-based data, or data collected by a harvester during operation can be used for updating past forest information; see e.g. Czaplewski (1990), Czaplewski & Thompson (2008), Ehlers et al. (2013).

Which methods to use and how frequently, is ultimately a question about added value that the more accurate, or less uncertain, information provides. Meanwhile, taking the maximum out of the data sets available requires combining or fusing of the information provided by the varying data sources (Khalegi et al. 2013).

However, even if we have access rights for a data set it may be located on some server at some physical location and may have a custom data structure format without a proper interface to access. Hence the many heterogeneous data sources available from a forest site can be difficult to access and approach by most others than the data provider itself.

Effective utilization of available forest resources is thus not only based on short-cycled, more-and-more accurate, even cost-effective data inventory methods. Instead, by providing easy access to best available up-to-date information on forests is expected to generate new applications and businesses and bring together varying users, e.g., sellers and buyers in wood trade, thus enhancing the utilization of forest resources.

In this paper we consider the definition and requirements for a forest data platform which acts as a service between data sources and applications by defining a common interface through which the varying heterogeneous data sources can be accessed and queried by applications; see Fig. 1. The data

sources can be physically distributed anywhere and the platform then collects the queried data from the actual data bases.

We will define a compact format for data and metadata in which all data sources are given to the platform and in which the platform hands out all the queried and possibly further processed data to the applications, thus enabling homogeneous view to data.
One particularly important piece of metadata considered in this paper is the uncertainty of the data and the estimates, enabling the fusion of varying data sources. Furthermore, in decision making uncertainty provides means for understanding the risks related to decision alternatives.

We define a grid cell as the basic unit in the data structure for all data and estimates. In principle, all the data needs to be either collected at or converted into this grid map. However, the data structure also supports definitions at any other forest area units, such as forest stands, or even definitions of individual trees. Hence, utilization of other formats for data definition is up to application implementations of the unit transformations.

To make the use of all available data simple for applications, we will define basic services provided by the platform for combining, or fusing, multiple data sources; see e.g. Khaleghi (2013). One basic service updates the data sources through growth prediction models (Pretzsch 2009, Burkhart & Tomé 2012) to the present time, or any other time in the future. Another service then applies Bayesian data fusion (Bishop 2006, Särkkä 2013, Khaleghi et al. 2013) for combining the data sources as an up-to-date fused estimate.

From an architectural perspective the basic services provided by the platform are simply applications that are replaceable by any other application. Hence there can be several implementations of growth prediction models for data updating purposes. Furthermore, it is by no means mandatory to use any such data updating or fusing services; instead, the platform can simply return queried data sets within the data structure format as such.

The rest of the paper is organized as follows. Section 2 discusses the forest data and considers the specification of data and metadata in general, and for uncertainty in particular. Platform design and the specification of the data structure are covered in Section 3. Section 4 discusses the general principles of data updating through prediction and Bayesian data fusion which further set some requirements for the data structure definition.

Section 5 covers an example case with two real forest data sets with the aim in testing the data structure and one possible implementation of the data updating and fusion methods. Section 6 concludes the paper.

2 FOREST DATA AND REQUIREMENTS FOR METADATA

Up-to-datedness, resolution and accuracy vary between forest data sets. For maximum utilization of heterogeneous data sets, the platform specifies a structure for representing data homogeneously.

Definition of a common data structure requires structures for representing the data itself, and in particular, the metadata which is the key for effective utilization of the data. Most important metadata includes the inventory date of the data enabling data updating through growth prediction, and uncertainty metadata enabling assessment of the data accuracy itself, but also the fusion of several data sets.

2.1 Forest data

Forest data is represented for individual trees, or for a collection of trees within a grid cell, or at some larger geographical area. The basic unit we are considering here is a 16mx16m grid cell on a square lattice. This is also the basic unit of LiDAR based forest inventory currently done in about 10 year cycles in Finland (Suomen metsäkeskus 2016). However, data given in other formats, such as on larger geographical areas or even as individual trees, are acceptable as well, but requires transformation to the grid cell units when used in analysis.

From raw data varying methods are typically applied to produce measures such as diameter (at breast-height, dbh) and height of individual trees, or average diameter and height of a collection of trees. Even distribution information can be given, e.g., defining the scale and shape parameters of the Weibull diameter distribution (Weibull 1951) and the parameters of the Näslund's height curve (Näslund 1936). We denote a parameter vector collecting all these varying attributes together at a grid cell by a $Nx1$ vector θ.

We should be able to apply the same growth prediction models to all the data sets, and being able to combine the up-to-date parameters with data fusion methods. However, varying inventory methods can produce varying attributes to θ. Furthermore, two data sets a and b producing θ_a and θ_b do not generally represent the same time instant. Therefore, we should first convert the vectors to representing similar information, and then update

them through growth prediction to represent the situation at the same time instant.

2.2 Internal data

As a representational format of data inside the platform we use joint distribution information of tree diameters (dbhs) and heights at the grid cells. In practice, we can use either a joint diameter-height distribution, or diameter distribution combined with a Näslund's height curve (Näslund 1936).

After transforming data into this format, θ consists of the distribution/height curve parameters. Common characteristics attributes, such as mean diameter and mean height, can then be re-calculated from the distribution information.

Because of their compactness in representation, we will consider parametric distributions, represented by a vector of distribution parameters. There exist methods to predict (Siipilehto 2006, 2011) or recover the diameter distribution (Siipilehto & Mehtätalo 2013) of the trees, or even the diameter-height joint distribution information from the characteristics attributes (Siipilehto 2011) defined at some geographical area.

It is not essential which method exactly to use, but rather that it enables the generation of diameter-height information on individual trees. Obviously, data already defined in a suitable format, does not require any transformations.

2.3 Forest metadata

Metadata should at least specify the inventory time, i.e. the date when the data has been collected, and information about the accuracy or uncertainty of the data. There can also be other metadata available, such as related to the inventory method used.

When the inventory time is known, it is possible to predict the data up-to-date at present time by utilizing a growth prediction model. Knowing uncertainty of data is itself valuable e.g. in decision making, but it also

enables the comparison of varying data sets and their combination through data fusion.

Uncertainty metadata is usually related to the data inventory method, but can also be affected e.g. by local weather conditions at the time of the inventory. Uncertainty can also vary between tree species or between different canopy layers.

2.4 Uncertainty description

Uncertainty related to an observation process can be obtained through reference data which is assumed to produce attribute values without measurement uncertainty. Uncertainty for scalar attribute θ_i as RMSE (root-mean-square error) values is calculated as

$$RMSE\left(\theta_i^{obs}\right) = \sqrt{E\left(\left(\theta_i^{obs} - \theta_i\right)^2\right)}, \quad (1)$$

where θ_i^{obs} is the value obtained (estimated) through observation and θ_i is the true (reference) value. For an unbiased estimator RMSE is equal to the standard deviation.

Uncertainty related to an observation process is typically described by means of conditional Normal distribution, yielding probabilities for observed values θ_i^{obs} given the reality θ_i. We consider a conditional multivariate Normal distribution as we are interested of correlations of the attribute estimation errors.

The observation model is here formulated as

$$\theta^{obs} = A\theta + B\mathrm{e}, \quad (2)$$

where θ is the true (reference) $Nx1$ parameter vector, θ^{obs} is the observed/estimated $Nx1$ parameter vector, and e is $Nx1$ vector specifies measurement uncertainty. A and B are NxN matrices.

Let us assume that only a single observation is obtained from each grid cell, and assume a set of grid cells with reference values $\left\{\theta_i^{ref}\right\}_{i=1}^{R}$ and observation-based estimates $\left\{\theta_i^{obs}\right\}_{i=1}^{R}$. The observation model is now

composed of observations obtained at varying grid cells. Hence, we first calculate the estimation errors for each cell as $\mathrm{e}_i = \theta_i^{obs} - \theta_i^{ref}$.

Estimating the mean μ_e and the covariance matrix $\mathrm{C_e}$ of the error vector, we end up with a multivariate Normal distribution $p_e(\mathrm{e}) \sim N(\mu_\mathrm{e}, \mathrm{C_e})$ of the estimation/observation error. Assuming A and B in the observation model as identity matrices, the observation model can then be written as

$$p\left(\theta_i^{obs} \middle| \theta_i\right) = p_e\left(\theta_i^{obs} - \theta_i\right), \quad (3)$$

where the probability of observation is proportional to the difference between the observed and the true parameter vector. Even though we described how the observation model can be obtained from reference data for all the grid cells, this does not mean that we could not use individual observation models even for each grid cell separately, if available.

As the uncertainty related to the observations can be affected by local conditions, the model could be written as $p\left(\theta_i^{obs} \middle| \theta_i, \alpha\right)$, where the parameter vector α specifies the dependency on the external conditions, such as soil type and degree days. However, this is not considered here.

3 DATA PLATFORM DESIGN

Platform design is described here at a rather high level, instead of technical details related to the implementation. The focus is on basic services provided by the platform, and on the specification of the data structure for data sources and estimates, together making the data sources easily accessible to any users and applications.

3.1 Architecture and interfaces

The platform operates between data sources and applications as depicted by Fig. 2. The platform acts as a data inquiry interface to the applications by providing a uniform view to the heterogeneous data sources. The data sets can be physically located anywhere and the platform gathers the inquired data from their physical locations.

The obvious advantage of such an interface is that there only needs to be one implementation to access each data source. Hence, instead of N applications each implementing its own accesses to M data sources, each application only needs to communicate with the inquiry interface of the platform.

The platform defines a data structure for exchanging data and metadata from applications to platform and from platform to applications. The data structure is designed common to all data sources, the up-to-date data estimates, and the fused data estimate. Hence the applications communicate with the platform only through the specified single data structure format.

As the goal is to make the data easily accessible to the users and applications, the platform implements the following three basic services.
The first simply returns inquired data sets as such to the application, represented in the format of the common data structure.
The second, additionally to the first, applies desired growth prediction model to the inquired data sets with a given prediction horizon, and then returns the data sets in the common data structure format.
The third, additionally to the second, applies Bayesian data fusion to the inquired up-to-date data sets, and returns also the fused estimate in the data structure format in addition to the updated data sets.

The platform can be implemented to follow service oriented architecture (SOA), where each application provides services to the others. For example, the basic services produce services to any other applications, and are thus also replaceable by any other application implementations. More generally, applications should provide services that can be exploited by other applications, thus encapsulating generally used methods as general service components.

The platform does not require implementing any physical data bases for storing data on its own, but instead can rely on the original data sources – hence the implementation phase becomes easier. Furthermore, were the platform implemented as a cloud application, there would be no need for setting up hardware for computational needs either, instead the cloud would provide very flexible and scalable framework for the platform implementation.

3.2 Data structure

There are some general design principles for the specification of the data structure. First of all, it is essential that the data structure is kept as compact as possible, since the aggregated amount of data from grid cells is very large when compared, e.g., to the definition at larger, and more traditional, stand level.

For the sake of compactness we also prefer parametric distribution representation over histograms. After all, the number of parameters required by a histogram representation is easily tenfold to the parametric distribution representation. Moreover, representing uncertainty information of histogram classes and their covariances becomes quite tricky.

The uncertainty related to observed parameter vector θ^{obs} is described compactly by the error covariance matrix (see Section 2), capturing the uncertainties and the error correlations. The same description can also be used for capturing the uncertainties of transformed parameters, as discussed in Section 3.

The data structure is depicted in Fig. 4. The output data from the platform can contain the updated input data sets and the fused estimate. These are given within the same data structure format as the input data sets. The data structure can be easily implemented, e.g., in JSON (JavaScript Object

Notation), or XML (eXtensible Markup Language) format. In the following we describe the basic structural constructs of the data structure.

General contains general metadata, e.g., related to the data inventory Method and the Date of inventory/observation. Type specifies if the data describes existing trees or removed/collected trees. Such a piece of information is needed, e.g., when updating existing information with information on a forest operation.

Object is a general construct for describing individual trees, grid cells, or some larger geographical area, such as a forest stand. Object contains field Sub objects which is an option to describe a list of objects and their probabilities of belonging to the particular Object's area.

Basic attributes contain attributes that are either categorical values or values calculated from other attribute values, such as development class, main tree species, or dominating height. The uncertainty of these attributes is not specified.

Attributes contains the description of each actual attribute within the Attribute construct. Basic parameters represent attributes which are typically discrete valued (or categorical) and with respect to which uncertain attributes are specified. Examples include tree species and storey class.

Parameters represents attributes in a vector form. For example, characteristics attributes, such as mean diameter, mean height, etc., can be given inside the vector. Weibull-distribution related parameters can be given similarly either in the same or in another Attribute construct. Also transformations of the parameters can be described.

Uncertainty defines the parameter uncertainties either through RMSE errors, as variances or as a covariance matrix – even for each grid cell separately. Uncertainties outside the Attributes construct can be used for defining covariances between the parameters defined at different Attribute constructs.

4 BASIC SERVICES FOR DATA UPDATING AND DATA FUSION

In this section we will consider the general principles of data estimate updating with prediction and combining of data estimates by applying Bayesian data fusion methods (Särkkä 2013). These are the basic services needed in general, and provided here to the applications by the platform.

In the literature, there is some discussion on applying Kalman filtering in estimating forest characteristics attributes for updating existing forest inventory data (Czaplewski 1990; Czaplewski and Thompson 2008; Ehlers et al. 2013). In particular, Ehlers et al. (2013) have applied Kalman filtering based methods to the plot-level attributes directly. We apply filtering to parameter vector containing diameter-height distribution related parameters either provided directly by data sources or by conversion from forest characteristics attributes.

4.1 Bayesian data fusion

Let's consider a grid cell i at time instant t. The general principle of data fusion is that of using observation $\theta_{i,t}^{obs}$ through observation model $p(\theta_{i,t}^{obs}|\theta_{i,t})$ to update some existing a priori information $p^{ap}(\theta_{i,t})$ to gain updated, or a posteriori, information on $\theta_{i,t}$ (Bishop 2006; Särkkä 2013):

$$p^{pos}(\theta_{i,t}|\theta_{i,t}^{obs}) \propto p(\theta_{i,t}^{obs}|\theta_{i,t})p^{ap}(\theta_{i,t}). \quad (4)$$

Here we assume that the observation model is independent on time instant t. However, t could be one of the parameters in α, discussed earlier.

Eq. 4 describes recursive updating of the information; new observation is instantly being used for updating the a priori information, see Fig. 3. The same update procedure can also be described in a batch mode, where we take all past observations $\{\theta_{i,t_k}^{obs,k}\}_k$, then predict each up-to-date to the present time to obtain $\{\theta_{i,t}^{obs,k}\}_t$, and finally obtain the fused estimate as (Särkkä 2013)

$$p^{pos}\left(\theta_{i,t}\middle|\{\theta_{i,t}^{obs,k}\}_t\right) \propto [\Pi_k p(\theta_{i,t}^{obs,k}|\theta_{i,t})]p^{ap}(\theta_{i,t}). \quad (5)$$

Here each $p\left(\theta_{i,t}^{obs,k}\middle|\theta_{i,t}\right)$ essentially defines the likelihood of the parameter vector given the observation. If no a priori information exists, we can simply drop the a priori term from Eq. 5.

Whether to use recursive or batch updating depends on the purpose. If we want to keep an up-to-date estimate and apply growth model to the single combined estimate, we go for recursive estimation. However, if we want to generate up-to-date combined estimate only when inquired and apply growth model individually to each observation, we choose batch updating.

As forest data are time series data obtained at varying time instants, the a priori information $p^{ap}(\theta_{i,t-1})$ needs to be predicted to time instant t to obtain $p^{ap}(\theta_{i,t})$. In principle, the growth prediction can be formulated as a state transition probability model $\pi(\theta_{i,t}|\theta_{i,t-1})$ to obtain the predicted parameter vector as

$$p^{ap}(\theta_{i,t}) = \int_{\theta_{i,t-1}} \pi\left(\theta_{i,t}\middle|\theta_{i,t-1}\right) p^{ap}(\theta_{i,t-1})\, d\theta_{i,t-1}. \qquad (6)$$

If the model makes annual predictions, and the prediction horizon is T years, the prediction step can be recursively repeated for T times.

4.2 About implementation

For the platform, we choose the batch updating because the platform performs data inquiries on-demand and based on the users' access rights to the data sources. Hence the data sets taken to data fusion can vary from user to user. Additionally, at least in principle, the platform does not keep-up a data base itself where recursively updated information could be stored.

There are many other ways to implement the prediction and the Bayesian data fusion. For Normal distributions and linear prediction models, the calculations have analytical solutions, and the computations are fast (Bishop 2006). It is also possible to linearize a non-linear prediction model or to approximate some other distribution with a Normal distribution.

Parameters of θ can also be transformed, e.g., by applying log transformation, so that in the transformed coordinates the use of Normal distribution is more justified; see e.g. Siipilehto (2011). Other options include sampling or particle based methods enabling rather accurate calculations, but they come with some additional computational costs (Särkkä 2013).

More generally, computational complexity of the data updating and fusing services depends much on their particular implementations. However, as we have considered grid cells independently on each other, the problem parallelizes, enabling massive parallel computations when necessary. However, this becomes a bit trickier if spatial dependencies (Fox et al. 2007a, 2007b) of the grid cells needs to be utilized, e.g., in individual growth prediction models.

It needs to be denoted, that data fusion requires that the data sets are not obtained through some earlier data fusion. In such a case it would be possible, e.g., to repeat the same data fusion again which would distort the results.

5 EXAMPLE CASE: FUSION OF TWO DATA SETS ON THE PLATFORM

This example demonstrates how two forest inventory data sets can be represented in the defined data structure format, and in particular, the principles of applying the updating and fusing methods to the real data. Instead of providing detailed results, the aim is in showing the basic principles of applying the methods to real data.

5.1 Inventory data

One of the data sets is based on satellite imaging and field measurements, while the other one is based on airborne laser measurements and field measurements. Both data sets are represented on the 16mx16m grid cells by the following set of average or aggregated characteristics attributes: mean diameter, mean height, total basal area, volume, and age. As only LiDAR based data separates these attributes for Norway spruce, Scots pine, and birch, we only consider all the species together.

The RMSE estimation errors of the characteristics attributes are given. Unfortunately, covariance or correlation information is not available. However, a correlation matrix of the estimation errors was calculated from a reference data and data obtained with the same LiDAR approach at another forest area. The correlations are expected to be similar as with the studied LiDAR data. Even though there is no guarantee that these correlations are close to those with the satellite-based data, we use them together with the known attribute RMSE errors to construct covariance matrix for the satellite based data.

Internal data representation and data fusion

For each data set and at each grid cell, the initial observed attributes, denoted by $\beta_{i,t}^{obs}$, together with the covariance matrix are first transformed into a joint distribution of Weibull diameter (dbh) distribution parameters (Weibull 1951) and Näslund's height curve parameters (Näslund 1936), denoted jointly by parameter vector $\theta_{i,t}$. The Näslund's height curve depicts the tree height as a function of tree diameter.

We follow the approach of Siipilehto (2011). Although it defines the models particularly for Scots pine, we apply the same models for all the tree species together. In the method, a basic prediction for both the parameters and the

attributes is first obtained based on the age attribute, and some soil-condition-related data (not available here). Then the basic prediction yielding $\theta_{i,t}$ is being calibrated with the observed attribute values $\beta_{i,t}^{obs}$ based on the conditional distribution $p(\theta_{i,t}|\beta_{i,t}^{obs})$. This yields the best linear unbiased predictor for the parameter vector, and utilizes covariances obtained when fitting the basic prediction model parameters; see details from (Siipilehto 2011).

As such this method does not consider the measurement uncertainty related to the observed attributes or age used by the basic prediction. Hence we actually need to consider

$$p(\theta_{i,t}|p(\beta_{i,t})) = \int_{\beta_{i,t}} p(\theta_{i,t}|\beta_{i,t})p(\beta_{i,t})d\beta_{i,t}, \quad (7)$$

which has an analytical solution if both of the distributions are (multivariate) Normal. The distribution $p(\beta_{i,t})$ can be either the likelihood obtained from the observation model $p(\beta_{i,t}^{obs}|\beta_{i,t})$ (no a priori information) or be the a posteriori distribution $p(\beta_{i,t}|\beta_{i,t}^{obs})$ where the a priori information can be the distribution yield by the basic prediction.

Linear growth model for $\theta_{i,t}$ together with Normal distributions, would result in analytically solvable final distribution $(\theta_{i,t'})$ $(t' > t)$. Such (non-linear) models exist at least for the Weibull distribution parameters (Pretzsch 2009), and non-linear models can be linearized to yield an analytical solution. This would be an appropriate solution for the rather short prediction horizons the updating procedure is typically handling.

However, we use a sampling based approach that is compatible with individual growth prediction models (Pretzsch 2009, Burkhart 2012) as they usually yield more accurate results. We first use the measurement model to sample instances from the joint distribution of the attributes together with the age attribute. Each sample is used to perform basic prediction with the sampled age attribute, and then the resulting parameter vector is calibrated with the sampled attribute values.

Calibrated parameter values are further used to sample diameter instances form the Weibull distribution. Each diameter sample is predicted up-to-date with a growth model. Finally, the predicted Weibull parameters together with the Näslund's height curve parameters are collected together (or their

transformed versions) and approximated by a multivariate Normal distribution, yielding the up-to-date distribution $p(\theta_{i,t'})$.

Having up-to-date distribution $p(\theta_{i,t'})$ for each data set, we apply the Bayesian data fusion of Eq. 5 to finally obtain the up-to-date fused estimate. The characteristic attributes and their uncertainty can then be re-calculated from the distribution. Also estimates at larger stand level can be calculated form the grid-based estimates, although not presented here.

5.2 Data structure

Both the input and output data sets are defined in the data structure format. For input data, measured attributes are represented by a parameter vector. Uncertainties are represented by covariance matrices. The prediction horizons are determined from given inventory dates.

In the returned fused estimate, the Weibull distribution parameters and the Näslund's height curve parameters are jointly represented by a parameter vector. Uncertainty of the parameter vector is represented by a covariance matrix. The calculated attribute values representing mean or aggregated values are represented by another vector together with a covariance matrix. The covariances between the distribution related parameters and the attributes are not presented, even though this is possible with the data structure.

5.3 Results

Although the approach is compatible with individual growth models generally, we use a very simple growth model here, based on an annual growth rate to make the data sets up-to-date at present time instant. Model adds annual Gaussian noise to the prediction, hence the uncertainty grows with the length of the prediction horizon.

There is approximately a one-year gap between the inventory dates of the two data sets, with the more uncertain satellite data being more recent. Although only a one difference, in this case the data fusion makes sense as we are updating older but less uncertain information with newer but more uncertain data.

We will illustrate the updating and data fusion only with one grid cell as the aim is only to give a conceptual idea of the use of the methods. Fig. 5 presents for the two data sets the marginal distributions for the up-to-date predicted Weibull parameters and the fused Weibull parameters. Also the respective Weibull distributions based on the mean values of the Weibull parameters are presented. Fig. 5 also presents the respective results for the Näslund's height curve parameters, and the Näslund's height curve corresponding to the mean parameter values.

The parameters of the LiDAR based data set are obviously less uncertain than the parameters of the satellite based data set. Hence the fused estimate follows more closely the LiDAR based estimates. However, the uncertainty of the fused estimates is smaller than the uncertainty of any of the data sets.

Fig. 6 represents an example of a stand-level dbh distribution aggregated from the Weibull distributions at the grid cells for a stand consisting of about 40 grid cells. In this case, we have first taken the Weibull distributions at the grid cells corresponding to the expectation values of the Weibull parameters. Then the stand-level distribution is obtained by taking a weighted sum of the distributions of the grid cells falling under the stand area, each weighted by its total stem count. Again the results are shown for the LiDAR data, satellite data, and the fused estimate. The fusion is done at the grid cells, as before. The results show how the stand-level distribution constructed from unimodal Weibull distributions can capture several modes. Again, the fused distribution falls somewhere between the LiDAR based and the satellite based distributions, but being closer to the less uncertain LiDAR based distribution. Each distribution is normalized with its expected total stem count (per hectare) at the grid cells. In this particular case, there is a rather large difference in the total stem counts between the satellite and the LiDAR based data.

5.4 Discussion

Deriving diameter distribution from each data set makes the methods compatible with individual tree prediction models (Pretzsch 2009). However, as the prediction time horizon is some years instead of decades as typically, e.g., in forestry planning (Rasinmäki 2007, Hynynen 2002), simpler and less computationally demanding models can be utilized. Such models are anyway preferable because of the use of the sampling based approach.

As the basic prediction (Siipilehto 2011) is based on stand age, this attribute should obviously be as accurate as possible, even though the method is rather robust if the calibration attributes are chosen correctly (Siipilehto 2011). In our data sets, however, the stand age variables are very uncertain, particularly in the satellite based data set. Hence in true applications utilizing this method with such data is rather questionable. Nevertheless, we have taken the uncertainty of stand age into account in the calculations, and the results look pretty decent.

There are some approximations done in the calculations. For example, the basic prediction model utilizes (mostly) logarithmic transformations which are then transformed back to arithmetic scale only after the calibration and data fusion. The transformed Weibull/Näslund parameters are also assumed multivariate normally distributed to make the final data fusion easier and analytically tractable, and in particular, to make the final representation compact for the data structure.

6 DISCUSSION

In this paper we considered a platform enabling uniform view to varying heterogeneous forest data sources through a common data inquiry interface and a specified data structure defining data, and in particular, the metadata related to the data sources. The platform offers basic services for data updating through prediction and combining of varying data sets through data fusion.

One of the key requirements of the data structure design is compactness of representation. Hence distribution information is preferably described by parametric distributions and the uncertainty metadata by either vectors of individual parameter uncertainties or by a covariance matrix of the estimation errors.

Although the data structure is here only defined for forest data, the future aim is to extend this specification to cover also other forest related data, such as soil conditions data. Soil type related data together with a short history of weather conditions, can be used to estimate traversability of the ground. This type of information is important for the annual planning and scheduling of forest operations and also for the route planning of harvester and other machines moving in the forest.

Spatial dependencies of the trees or grid cells or other units were not considered here. However, particularly as using smaller units for data and information representation, such as grid cells, modelling spatial dependencies becomes more essential. For example, growth models typically require information related to the surrounding environment to specify the competition related parameters.

The idea of the platform is in making access and utilization of data easier by abstracting the data into a source appearing uniform and by providing basic services for data updating and fusion. This is expected to essentially reduce the threshold for implementing new applications build on top of the platform, thus promoting forestry related businesses such as wood trade, forestry planning and worksite planning.

REFERENCES

Antropov, O., Rauste, Y., Ahola, H. & Hame, T. 2013. Stand-level stem volume of boreal forests from spaceborne SAR imagery at L-band. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 6: 35-44.

Bishop, C. M. 2006. Pattern recognition and machine learning, Springer.

Burkhart, H. E. & Tomé, M. 2012. Modeling forest trees and stands. Springer Science & Business Media.

Czaplewski, R.L. 1990. Kalman filter to update forest cover estimates. State-of-the-art methodology of forest inventory. Edited by V.J. Labau and T. Cunia. USDA For. Serv., Pacific Northwest Research Station. General Technical Reports 263: 457–465.

Czaplewski, R. L., & Thompson, M. T. 2008. Opportunities to improve monitoring of temporal trends with FIA panel data. In: Forest & Analysis Symposium: 21–23.

Ehlers, S., Grafström, A., Nyström, K., Olsson, H. & Ståhl, G. 2013. Data assimilation in stand-level forest inventories. Canadian Journal of Forest Research 43: 1104–1113.

Fox, J. C., Bi, H. & Ades, P. K. 2007a. Spatial dependence and individual-tree growth models: I. Characterising spatial dependence. Forest Ecology and Management 245: 10–19.

Fox, J. C., Bi, H. & Ades, P. K. 2007b. Spatial dependence and individual-tree growth models: II. Modelling spatial dependence. Forest Ecology and Management 245: 20–30.

Holopainen, M., Haapanen, R., Karjalainen, M., Vastaranta, M., Hyyppä, J., Yu, X. & Hyyppä, H. 2010. Comparing accuracy of airborne laser scanning and TerraSAR-X radar images in the estimation of plot-level forest variables. Remote Sensing 2: 432–445.

Hynynen, J. 2002. Models for predicting stand development in MELA system. METLA Vantaa Research Center.

Kankare, V., Vauhkonen, J., Tanhuanpää, T., Holopainen, M., Vastaranta, M., Joensuu, M. & Viitala, R. 2014. Accuracy in estimation of timber assortments and stem distribution – A comparison of airborne and terrestrial laser scanning techniques. ISPRS Journal of Photogrammetry and Remote Sensing 97: 89–97.

Khaleghi, B., Khamis, A., Karray, F. O. & Razavi, S. N. 2013. Multisensor data fusion: A review of the state-of-the-art. Information Fusion 14: 28–44.

Näslund, M. 1936. Skogsförsöksanstaltens gallringsförsök i tallskog. Meddelanden från Statens Skogsförsöksanstalt 28.

Pretzsch, H. 2009. Forest dynamics, growth, and yield. Springer Berlin Heidelberg. p. 1–39.

Rasinmäki, J. 2007. Management of multi-scale forest resource data over time. Dissertationes Forestales.

Raumonen, P., Kaasalainen, M., Åkerblom, M., Kaasalainen, S., Kaartinen, H., Vastaranta, M. & Lewis, P. 2013. Fast automatic precision tree models from terrestrial laser scanner data. Remote Sensing 5: 491–520.

Siipilehto, J. 2006. Linear prediction application for modelling the relationships between a large number of stand characteristics of Norway spruce stands. Silva Fennica 40: 517.

Siipilehto, J. 2011. Local prediction of stand structure using linear prediction theory in Scots pine-dominated stands in Finland. Silva Fennica 45: 4.

Siipilehto, J. & Mehtätalo, L. 2013. Parameter recovery vs. parameter prediction for the Weibull distribution validated for Scots pine stands in Finland. Silva Fennica 47: 1–22.

Suomen metsäkeskus. Metsätiedon keruu. Saatavilla: http://www.metsakeskus.fi/metsatiedon-keruu#. [Viitattu 20.12.2016].

Särkkä, S. 2013. Bayesian filtering and smoothing. Cambridge University Press.

Vastaranta, M., Niemi, M., Karjalainen, M., Peuhkurinen, J., Kankare, V., Hyyppä, J. & Holopainen, M. 2014. Prediction of forest stand attributes using TerraSAR-X stereo imagery. Remote Sensing 6: 3227–3246.

Weibull, W. 1951. Wide applicability. Journal of applied mechanics 103: 293–297.

ANNEXES

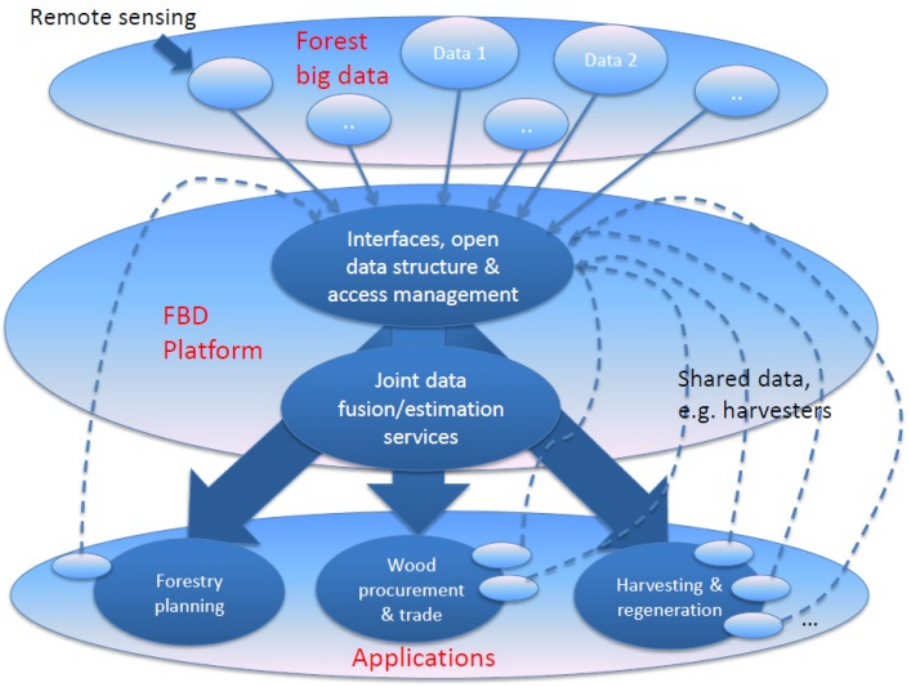

Fig. 1. The general idea of the data platform. Platform functions as a data inquiry interface between data sources and applications.

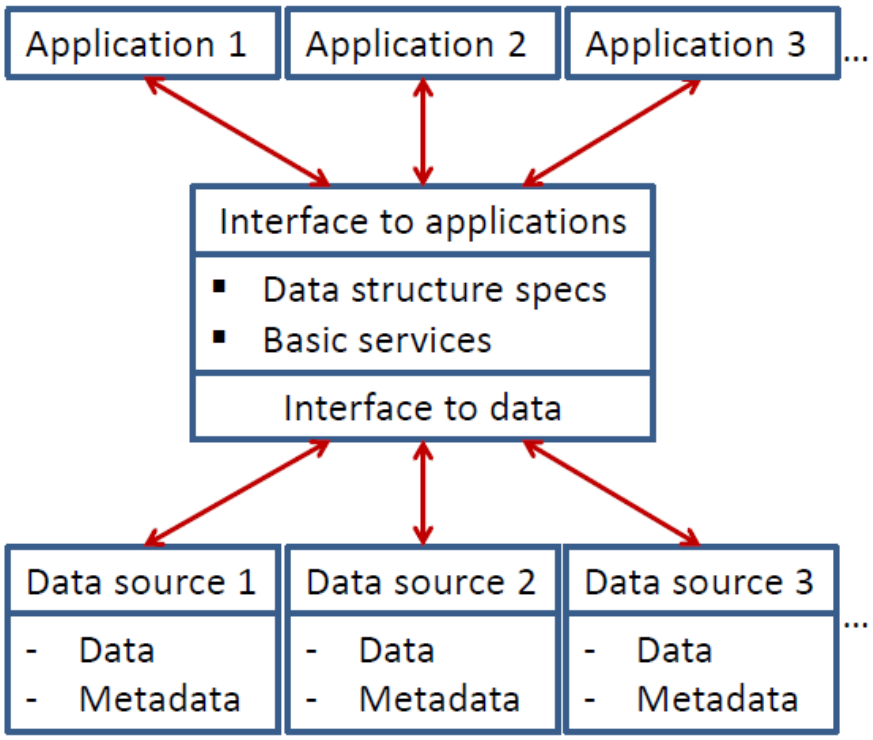

Fig. 2. Simplified architectural description of the platform.

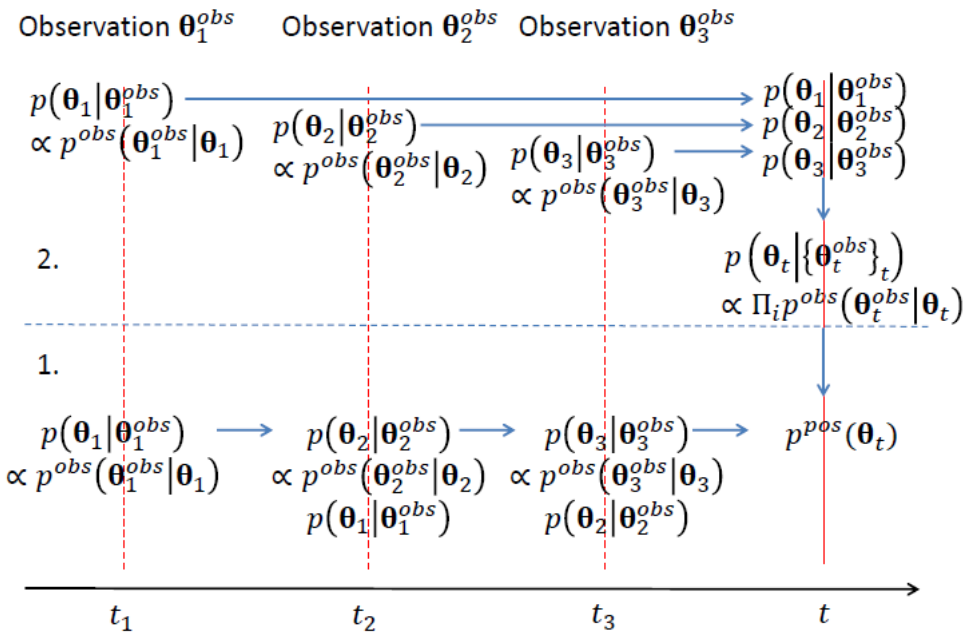

Fig. 3. Data fusion alternatives. Data fusion can be done either as 1) recursive estimation or 2) as batch estimation.

General

- **Date**: 3x1 integer
- **ID**: integer
- **Owner**: string
- **Type**: integer
- **Description**: string
- **Method**: integer
- **Area**: Nx2 real

Objects

- **Object**
  - **Date**: 3x1 integer
  - **ID**: integer
  - **Type**: integer
  - **Description**: string
  - **DataSourceIDs**: Nx1 integer
  - **Location**
    - **Value**: Nx(2-3) real
    - **Uncertainty**: 2(-3)x2(-3) real
    - **ExistenceProbability**: real
  - **SubObjects**
    - **Type**: Nx1 integer
    - **Description**: Nx1 string
    - **IDs**: Nx1 integer
    - **MemberProbability**: Nx1 real
  - **BasicAttributes**
    - **Type**: Nx1 integer
    - **Description**: Nx1 string
    - **Values**: Nx1 integer
    - **Probability**: Nx1 real
  - **Attributes**
    - **Attribute**
      - **BasicParameters**
        - **Type**: Nx1 integer
        - **Description**: Nx1 string
        - **Values**: Nx1 integer
        - **Probability**: Nx1 real
      - **Type**: integer
      - **Description**: string
      - **Parameters**
        - **Type**: Nx1 integer
        - **Description**: Nx1 string
        - **Values**: NxM real
        - **NormalizationID**: integer
      - **Uncertainty**
        - **Type**: integer
        - **Description**: string
        - **Transformation**: Nx2 integer
        - **Values**: Nx1, Nx2, or NxN
    - …
  - **Uncertainties**
    - **Uncertainty**
      - **BasicParameters**: Nx2 integer
      - **Types**: Nx2 (integer, integer) pairs
      - **Transformation**: Nx2 integer
      - **Values**: Nx1 vector
    - …
- …

Fig. 4. Data structure specification.

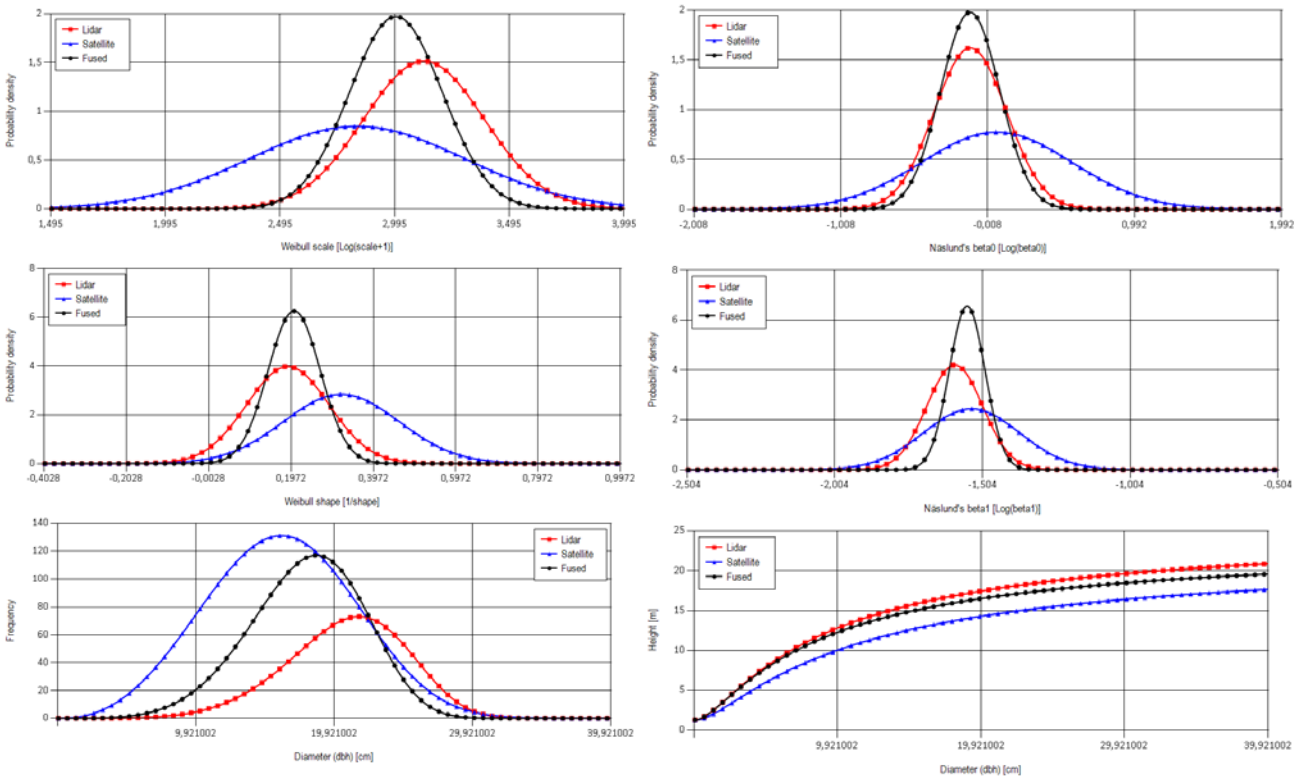

Fig. 5. Estimation results of the transformed Weibull distribution parameters and Näslund's height curve parameters. Left column: marginal distributions of Weibull scale (top) and shape (middle) parameters, and Weibull distribution calculated from the mean values of the parameters (bottom). Right column: marginal distribution of the Näslund's height curve β_0 (top) and β_1 (middle) parameters, and Näslund's height curve calculated from the mean values of the parameters (bottom).

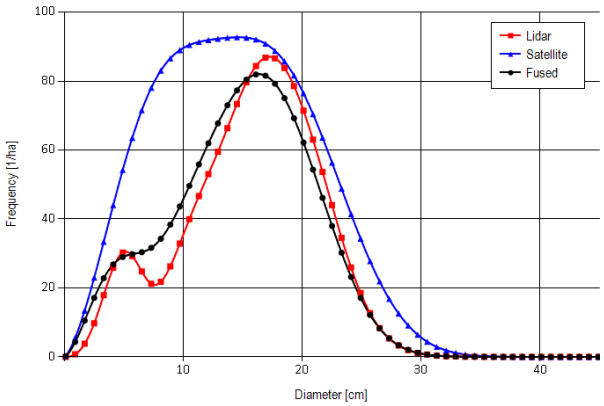

Fig. 6. Example of stand distribution obtained by taking a weighted sum of the Weibull distributions at the grid cells. Each grid cell is weighted by the total stem count at the grid cell.