Atmos. Meas. Tech. Discuss., 7, 7137–7174, 2014
www.atmos-meas-tech-discuss.net/7/7137/2014/
doi:10.5194/amtd-7-7137-2014

Atmospheric Measurement Techniques Discussions
Open Access

Regression models tolerant to massively missing data: a case study in solar radiation nowcasting

I. Žliobaitė[1,2], **J. Hollmén**[1,2], **and H. Junninen**[3]

[1]Aalto University, Department of Information and Computer Science, Espoo, Finland
[2]Helsinki Institute for Information Technology (HIIT), Helsinki, Finland
[3]Department of Physics, University of Helsinki, Helsinki, Finland

Received: 14 April 2014 – Accepted: 24 June 2014 – Published: 16 July 2014

Correspondence to: I. Žliobaitė (indre.zliobaite@aalto.fi)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Statistical models for environmental monitoring strongly rely on automatic data acquisition systems, using various physical sensors. Often, sensor readings are missing for extended periods of time while model outputs need to be continuously available in real time. With a case study in solar radiation nowcasting, we investigate how to deal with massively missing data (around 50 % of the time some data are unavailable) in such situations. Our goal is to analyze the characteristics of missing data and recommend a strategy for deploying regression models, which would be robust to missing data in situations, where data are massively missing. We are after one model that performs well at all times, with and without data gaps. Due to the need to provide instantaneous outputs with minimum energy consumption for computing in the data streaming setting, we dismiss computationally demanding data imputation methods, and resort to a simple mean replacement. We use an established strategy for comparing different regression models, with the possibility of determining how many missing sensor readings can be tolerated before model outputs become obsolete. We experimentally analyze accuracies and robustness to missing data of seven linear regression models and recommend using regularized PCA regression. We recommend using our established guideline in training regression models, which themselves are robust to missing data.

1 Introduction

Environmental monitoring strongly relies on automatic data acquisition systems, using various physical sensors. For instance, SMEAR stations[1] measure the relationship of atmosphere and forest in boreal climate zone (Hari and Kulmala, 2005). They are equipped with an extensive range of measurement interests: atmospheric and flux measurements, irradiation and flux measurements, tree physiology measurements, soil

[1] http://www.atm.helsinki.fi/SMEAR/

AMTD
7, 7137–7174, 2014

and soil-water measurements, and solar irradiance. Due to the continuous flux of measurements, the setup can be put to the framework of streaming data. Statistical models built on such streaming data, see e.g. (Hrust et al., 2009; Lu et al., 2006; Menut and Bessagnet, 2010), need to operate continuously and provide outputs in real time.

Physical sensors are exposed to various risks due to severe environmental conditions, exposure to physical damage or battery drainage. Under such circumstances it is very common to have time intervals when data from some of the sensor readings are missing. A lot of advanced missing value imputation schemes have been developed (Junninen et al., 2004; Allison, 2001) primarily targeting offline exploratory data analysis, where computational resources are practically unlimited and it is critical to reconstruct data as accurately as possible. A simple mean replacement remains popular in regression modeling (Kadlec et al., 2009), in situations where real time outputs are needed, computational resources and time are limited, but the input data does not need to be reconstructed perfectly accurately as long as model outputs remain correct.

The goal of this study is to experimentally analyze the performance and robustness of linear regression models to massively missing data for operation in resource aware settings. We consider the situations where data are massively missing, which means that around 50 % of the time at least one sensor does not deliver readings and there is no single sensor that dominates the missing data, data from any sensor can be missing. In such a situation, readings from input sensors may be missing for extended periods of time, but nevertheless, model outputs need to be produced continuously and delivered in real time, withholding from making model outputs when some data are missing is not an option. We aim at building one regression model that is robust in performance, i.e. the expected performance is stable no matter how many sensor readings are missing.

We present a case study in solar radiation nowcasting using meteorological sensor data as inputs, where multiple sensor failures happen frequently due to environmental and operational reasons. We analyze the performance of seven linear regression

AMTD
7, 7137–7174, 2014

models coupled with mean replacement of missing values, and provide recommendations for robust and accurate modeling in such circumstances.

The paper presents a case study, where our earlier published results (Žliobaitė and Hollmén, 2013), are put to practice in a solar radiation nowcasting problem in the context of a SMEAR measurement station (Hari and Kulmala, 2005). The reader interested in the theoretical underpinnings of our approach and a follow up is advised to read studies (Žliobaitė and Hollmén, 2013, 2014), whereas the current paper focuses on practical implications of the results, and demonstrates how regression problems with lots of missing data can be successfully solved with our recommended scheme. The results apply to the case of linear regression coupled with mean replacement of missing values.

Research attention to solar radiation nowcasting and short term forecasting using statistical data driven models is increasing due to growing popularity of solar energy plans, that need forecasts for planning. Research studies mostly focus on searching for a suitable statistical modeling technique: artificial neural networks (Marquez and Coimbra, 2011), autoregressive time series models (Bacher et al., 2009), Markov models (Bhardwaj et al., 2013), or optimally integrating different data sources, such meteorological variables, ground and remote sensing observations, or satellite images (Hammer et al., 1999; Vuilleumier et al., 2011). We are not aware of any research work addressing the problem of massively missing values in solar radiation forecasting.

The rest of the paper is organized as follows. Section 2 describes the SMEAR data used in the case study, the methodology of the modeling, and the experimental protocol. Section 3 presents and discusses the results of the case study. Section 4, summarizes the contributions and concludes the study.

AMTD
7, 7137–7174, 2014

2 Materials and methods

2.1 Data

We use a data stream recorded at SMEAR II station in Hyytiälä, Finland (Junninen et al., 2009) (61°50′51″ N, 24°17′41″ E, 181 m a.s.l.), measuring the forest ecosystem–atmosphere relationships. We use data over seven years period (April 2005–April 2013), recorded every 30 min from 37 observation sensors. The data coming from the station has on average 7 % of missing values. Missing values may occur due to occasional failure of measuring sensors, wear and tear, or variations in electricity power supply. At least some data are missing 50 % of the time. There is no single sensor that would provide non interrupted readings over those five years; for any sensor from 1 % (about 4 days per year) up to 25 % (3 months per year) values are missing.

The task is to nowcast the current level of solar radiation from the meteorological sensor data given in Table 1. Incoming radiation to the earth is constant (with the accuracy we care about) in a given time of the year and hour. The only unknown is the absorption in atmosphere and more importantly in the clouds and anthropogenic pollution plumes. Hence, an interesting variable to infer is the cloudiness, or, in other words, the deviation of the measured radiation from the theoretical maximum. In this schema, other meteorological parameters could be used to estimate the cloudiness and this can further used to calculate the actual radiation, but the primary variable to nowcast is the difference between the theoretical and actual radiation.

This nowcasting task would be relevant to the stations where no radiation measures are available. The station SMEAR II, from which the input data originate, can measure solar radiation; hence, the true values are available for us for evaluation purposes. However, instrumentation for measuring solar radiation is not always available. Small meteorological observation stations may not be able to afford to have solar radiation measured, but it may be interesting to nowcast radiation from meteorological data that is available anyway.

AMTD
7, 7137–7174, 2014

We define the target variable as the ratio: the actual radiation divided by the theoretical maximum radiation. This gives a value between 0 and 100 %, where 100 % indicates that all the theoretically possible radiation is actually incoming. The sensor Global RADIATION (Table 1) is not used as model input, it is only used for evaluating the nowcasting accuracy. It indicates the actual radiation and is used in forming the target variable.

The theoretical maximum radiation is calculated using MIDC SOLPOS Calculator[2]. SOLPOS is a computational tool that calculates the apparent solar position and intensity (theoretical maximum solar energy) based on the date, time, and location on Earth. The tool is developed (written in C) and maintained by The National Renewable Energy Laboratory, which is operated for the US Department of Energy by the Alliance for Sustainable Energy. The calculations are based on established models for solar position reported in (Michalsky, 1988) and other sources.

The following input parameters were used: Lat: 61.8475, Lon: 24.29472, time zone: 2 (location parameters), surface pressure 990 mbar, ambient dry-bulb temperature 3 °C, azimuth of panel surface 180°, degrees tilt from horizontal of panel 0, solar irradiance constant 1360.8 $W\,m^{-2}$ (Kopp and Lean, 2011), shadow-band width 7.6 cm, shadow-band radius 31.7 cm, shadow-band sky factor 0.04, interval of a measurement period 0 s.

The sensor readings are often correlated with each other. Figure 1 visualizes the pairwise correlations computed over non-missing data. We see distinct blocks of positive and negative correlations. For instance, relative humidity (RH) is negatively correlated with temperature (T).

[2] http://www.nrel.gov/midc/solpos/solpos.html

AMTD
7, 7137–7174, 2014

2.2 Prerequisites

2.2.1 Setting

Suppose we have r sources generating streaming data (e.g. weather observation sensors). Data are recorded in multidimensional vectors $\boldsymbol{x} \in \mathbb{R}^r$. Our task is to nowcast the target variable $y \in \mathbb{R}^1$ (e.g. solar radiation) from these sensor readings as inputs. The regression model is then $y = f(\boldsymbol{x}) = f(x_1, \ldots, x_r)$, and the corresponding learning task is to learn the approximate the function f from the available input-output data. It is important to note that we do not make use of temporal information of the variables, that is, we predict the value of the output y at time t, with the sensor readings available at the same time point t, hence the task is referred to as nowcasting. With the time index, the regression model is $y^{(t)} = f(x_1{}^{(t)}, \ldots, x_r{}^{(t)})$. In the rest of the paper, we omit the time index t. For the identities of the sensors used in the case study ($r = 36$), see Table 1.

Data arrive in real time, and nowcasting needs to be delivered as soon as possible, in nearly real time. The nowcasting performance should be stable in a sense that the expected loss in accuracy due to possible missing values should be minimal. Having in mind that often environment monitoring sensors are operating on batteries, or autonomous power sources, consumed computational resources for data processing, including missing value imputation, should be minimal.

2.2.2 Imputation of missing data

We assume that when a sensor fails, the missing values are automatically replaced with the *mean* value, which remains to be a popular approach in practice due to its simplicity and low user cost (Black et al., 2007; Kadlec et al., 2009; Enders, 2010). To keep the focus, we also assume that there is no need for any king of data driven missing value detection, the system knows when a value is missing.

In this study, we do not consider alternative imputation methods due to two reasons. Firstly, our main goal is to investigate the robustness of regression models to missing

AMTD
7, 7137–7174, 2014

data rather than select the best imputation scheme. Secondly, advanced model based imputation methods such as linear interpolation, nearest neighbor imputation, self organizing map or multilayer perceptron methods (Junninen et al., 2004) typically are more accurate when the amount of missing data is small, but lose their advantage when long missing data gaps are expected; while multiple imputation methods (Junninen et al., 2004) bear relatively high computational costs and are favorable in once-off imputation operations, but are not very suitable for continuous online operations and imputation in real time. More importantly, such methods implicitly or explicitly assume that data are missing at random, i.e. a sensor value being missing is independent both of observable variables and of unobservable parameters of interest. In reality this assumption may often be violated, for instance, sensors going off at low temperatures.

One could make imputation models using knowledge about physical relationships between variables. However, when a lot of data is missing, such approach would encounter a combinatorial explosion. There would need to be a model for each combination of missing variables, that means building and maintaining 2^r models, where r is the number of input features.

2.2.3 Performance indicators

We use a nowcasting error as the main measure of the performance, which is computed on a subset of data that was not used for parameter estimation (Hastie et al., 2001). The mean squared error (MSE) is a popular measure to quantify the discrepancy between the true target value y and the value output by the model $\hat{y}$. MSE punishes large deviations from the true values that is relevant in environmental monitoring applications, where large errors are to be avoided. For practical interpretability RMSE is often used, which is the square root of MSE, it reports the error in the same units as the target variable. For a test dataset MSE and RMSE are computed as

$$\mathrm{MSE} = \frac{1}{n}\sum_{l=1}^{n}(\hat{y}^{(l)} - y^{(l)})^2, \mathrm{RMSE} = \sqrt{\mathrm{MSE}}, \quad (1)$$

AMTD
7, 7137–7174, 2014

where $y^{(l)}$ is the true target value of the lth sample and $\hat{y}^{(l)}$ is the corresponding model output, n is the number of samples in the test set. In the experiments we report RMSE, which can be interpreted as an average deviation of model outputs from the true target values.

2.3 Computational methods

2.3.1 Linear regression model

For nowcasting, we consider linear regression models, which assume that the relationship between r input variables $\boldsymbol{x} = (x_1, \ldots, x_r)$ and the target variable y is linear. Without loss of generality we assume that the input data are standardized before modeling to have zero mean and unit variance[3]. The regression model takes the form

$$y = b_1 x_1 + b_2 x_2 + \ldots + b_r x_r + \epsilon = \boldsymbol{x}\boldsymbol{\beta} + \epsilon \quad (2)$$

where ϵ is the error variable and the vector $\boldsymbol{\beta} = (b_0, b_1, b_2, \ldots, b_r)^{\mathrm{T}}$ contains the parameters of the linear model (regression coefficients). Since the data are assumed to have been standardized there is no bias term in the model. In matrix form, the model is $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, where $\mathbf{X}_{n \times r}$ is a sample data matrix containing n records from r sensors, and $\boldsymbol{y}_{n \times 1}$ is a vector of the corresponding n target values.

2.4 Ordinary least squares

There are different ways to estimate the regression parameters (Hastie et al., 2001). Ordinary least squares (OLS) is a simple and probably the most common estimator. It

[3]For standardization we need to estimate the data mean m and the standard deviation s from a sample dataset, then $x_{\text{standardized}} = (x - m)/s$. For every variable we need to store the values m and s and to apply the same procedure to the new incoming data before nowcasting.

AMTD
7, 7137–7174, 2014

minimizes the sum of squared residuals giving the following solution

$$\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \arg\min_{\boldsymbol{\beta}} \left((\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) \right) = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}, \tag{3}$$

Having estimated a regression model $\hat{\boldsymbol{\beta}}$ nowcasting on new data $\boldsymbol{x}_{\mathrm{new}}$ can be made as

$$\hat{y} = \boldsymbol{x}_{\mathrm{new}}\hat{\boldsymbol{\beta}}. \tag{4}$$

2.4.1 Regularization

If the input variables are correlated with each other, the optimization problem could result in poor estimates for the parameters In such situations, regularization is often used for estimating the regression parameters. The Ridge regression (RR) (Hoerl and Kennard, 1970; Hastie et al., 2001) regularizes the regression coefficients by imposing a penalty on their magnitude. RR solution minimizes the following cost function

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\mathrm{RR}} &= \arg\min_{\boldsymbol{\beta}} \left((\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \right) \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}\boldsymbol{y}, \end{aligned} \tag{5}$$

where $\lambda > 0$ controls the amount of shrinkage: the larger the value of λ, the greater the amount of shrinkage. $\mathbf{X}$ denotes the $n \times r$ training dataset and $\boldsymbol{y}$ is the $n \times 1$ vector of the true target values, $\mathbf{I}$ is the $r \times r$ identity matrix. Nowcasting on new data $\boldsymbol{x}_{\mathrm{new}}$ can be made as

$$\hat{y} = \boldsymbol{x}_{\mathrm{new}}\hat{\boldsymbol{\beta}}_{\mathrm{RR}}. \tag{6}$$

2.4.2 Principal component regression

Principal component (PCA) regression (Jolliffe, 2002) first transforms the input data by rotating it towards its principal components and then estimates the regression coefficients on the transformed data.

AMTD
7, 7137–7174, 2014

Let $\mathbf{X}_{n\times r}$ be the training data matrix and $\mathbf{R}_{r\times k}$ is the matrix of k principal components, corresponding to the largest eigenvalues. Here k is a user defined parameter such that $1 \le k \le r$, if $k = r$ then PCA regression becomes the ordinary regression. Then OLS gives the following solution on the transformed input data

$$\hat{\boldsymbol{\beta}}^{\star}_{\mathrm{PCA}} = \arg\min_{\boldsymbol{\beta}} \left((\boldsymbol{y} - \mathbf{X}\mathbf{R}\boldsymbol{\beta}^{\star})^{\mathrm{T}} (\boldsymbol{y} - \mathbf{X}\mathbf{R}\boldsymbol{\beta}^{\star}) \right), \tag{7}$$

and in the original data space the solution is
$\hat{\boldsymbol{\beta}}_{\mathrm{PCA}} = \mathbf{R}\hat{\boldsymbol{\beta}}^{\star}_{\mathrm{PCA}}$. Nowcasting on new data $\boldsymbol{x}_{\mathrm{new}}$ can be made as

$$\hat{y} = \boldsymbol{x}_{\mathrm{new}}\mathbf{R}\hat{\boldsymbol{\beta}}^{\star}_{\mathrm{PCA}} = \boldsymbol{x}_{\mathrm{new}}\hat{\boldsymbol{\beta}}_{\mathrm{PCA}}. \tag{8}$$

2.4.3 Partial least squares regression

Partial least squares (PLS) regression is very popular in chemometrics (Wold et al., 2001). Similarly to PCA, the input data are transformed, but instead of maximizing the variance of the input data (as in PCA) this transformation maximizes the covariance between input variables and the target. There is no convenient analytical solution for optimization, instead an iterative optimization is employed for parameter estimation. The procedure is presented in Algorithm 1. Here k is a user defined parameter such that $1 \le k \le r$, if $k = r$ then PLS regression becomes the ordinary regression.

Nowcasting on new data $\boldsymbol{x}_{\mathrm{new}}$ can be made as

$$\hat{y} = \boldsymbol{x}_{\mathrm{new}}\hat{\boldsymbol{\beta}}_{\mathrm{PLS}}. \tag{9}$$

2.5 Estimating robustness of linear regression models to missing data

For linear regression models it is possible to determine theoretically how many missing inputs can be tolerated before model outputs become obsolete. We can estimate robustness of a linear regression model to potentially missing input data using the deterioration index (Žliobaitė and Hollmén, 2013), which is defined as

$$d = -\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\Sigma} - \mathbf{I})\boldsymbol{\beta}, \tag{10}$$

AMTD
7, 7137–7174, 2014

where $\boldsymbol{\beta}$ is a vector of the regression coefficients, assuming that the input variables have been standardized to zero mean and unit standard deviation, $\boldsymbol{\Sigma}$ is the covariance matrix of the input data and $\mathbf{I}$ is the identity matrix. High values of the index d indicate low tolerance to missing data. The prediction errors will increase fast with the number of missing inputs. The smaller d, the more robust to missing data the model is. d can be negative, that is the best option.

Low d guarantees robustness to missing data, but the models with low d do not necessarily give good predictions when all the data are available. Hence, a tradeoff between accuracy and robustness needs to be found, the following method can help to find it.

Suppose we get two models A and B, and we would like to select one for deployment. We can measure their prediction errors on a training dataset using cross-validation, $\mathrm{RMSE}^{(\mathrm{A})}$ and $\mathrm{RMSE}^{(\mathrm{B})}$ respectively. We can also compute deterioration indices $d^{(\mathrm{A})}$ and $d^{(\mathrm{B})}$. Without loss of generality assume that $\mathrm{RMSE}^{(\mathrm{A})} \geq \mathrm{RMSE}^{(\mathrm{B})}$, i.e. model B shows a better prediction accuracy when no data are missing. If $d^{(\mathrm{A})} \geq d^{(\mathrm{B})}$, then model B is also more robust. In such a case model B is better (or at least as good) in both characteristics, and hence B is preferred over A.

If, however, $d^{(\mathrm{A})} < d^{(\mathrm{B})}$, then we can find, how many input readings can go missing before A becomes better than B. The number $m^{\star}$ can be computed as (Žliobaitė and Hollmén, 2013)

$$m^{\star} = (r-1)\frac{[\mathrm{RMSE}^{(\mathrm{A})}]^2 - [\mathrm{RMSE}^{(\mathrm{B})}]^2}{d^{(\mathrm{B})} - d^{(\mathrm{A})}}, \tag{11}$$

where r is the total number of input sensors.

AMTD
7, 7137–7174, 2014

2.6 Experimental protocol

2.6.1 Data preparation and preprocessing

Solar radiation readings (target variable) are available 99 % of the time. We eliminate from the experiment the samples where no target value is available, since we cannot use such samples neither for model training, nor can we measure the model accuracy on them.

The following pre-processing of the target values is performed. If the measured solar radiation is negative, it is set to be zero. If the measured solar radiation exceeds the theoretical (maximum) radiation, the measurement is corrected to be equal to the theoretical radiation. In practice, this can happen when during a cloudy day the sky is clear where sun is shining, but a cloud cover is elsewhere. The cloud reflects back more back-reflected radiation that the blue sky. For simplicity, at this stage we do not consider this effect in this modeling.

Exploratory analysis of missing data is performed on the full dataset. For analysis the model accuracies we use the first *three* years of data as a training set and the remaining *four* years as the testing set. We assume the scenario where an analyst is currently at the end of year three, and all the previous three years of data are available for model calibration. After modeling and calibration is done, an online operation scenario is assumed, where the testing data (four years) arrive in the sequential order.

From the training set we eliminate all the observations that contain any missing values (34 % of train data). The testing set contains all samples no matter whether some values in the input data are missing. In addition, we eliminate from the training and testing sets all the observations where the value of theoretical radiation is 0 (the periods of dark), since then the value of the target variable is also 0, which can be nowcasted with 100 % accuracy, while when performing experimental comparison of models we are interested in accuracies of non-trivial nowcasting tasks.

The training data are standardized to have zero mean and unit standard deviation. The testing data are preprocessed by subtracting the mean and dividing by the

AMTD
7, 7137–7174, 2014

standard deviation calculated on the training set. After standardization we replace all the missing values in the testing set by zeros and test the regression models.

2.6.2 Regression models used in the experiments

We experimentally analyze seven regression models summarized in Table 2.

ALL uses all r sensors as inputs. SEL selects k sensors that have the largest absolute correlation with the target variable (correlation is measured on the training data) and builds a regression model on those k sensors. PCA rotates the input data using principal component analysis, k features corresponding to the largest eigenvalues are retained, and then builds a regression model on those k new features. PLS rotates input data to maximize the covariance between the inputs and the target. We keep k new features.

ALL, SEL and PCA use the ordinary least squares optimization procedure (OLS) for parameter estimation. In addition, we test the same approaches but using the regularized Ridge regression (RR), these models are denoted as ALLr, SELr and PCAr. PLS uses its own iterative optimization procedure, which is not regularized.

In addition, we compare the performance to a naive baseline NAI, which makes a constant output that the radiation will be the same as the average radiation in the training data.

2.6.3 Software and hardware

The experiments are performed in MATLAB 2012b using in-house produced code (no extra packages are required) on a commodity laptop computer (Processor 2.5 GHz Intel Core i5; Memory 8 GB 1600 MHz DDR3). The dataset used in this study and the code for the experiments are available[4] for research purposes.

[4] http://users.ics.aalto.fi/indre/smear.zip

AMTD
7, 7137–7174, 2014

3 Results and discussion

3.1 Analysis of missing data characteristics

We first analyze how missing values occur in the case study dataset. Figure 2a presents the distribution of missing sensors. We see that about half of the time nothing is missing, and half of the time observation vectors are incomplete. Over 35 % of the time 2–4 sensors are missing. The mean number of missing sensors over all the dataset is 2.4. We observe from the data that up to 36 sensors (that is all the input sensors) may be missing at a time. From this analysis we conclude that the amount of missing data is at a massive scale and scope, and missing values needs to be taken into consideration when building nowcasting models on this data. This also demonstrates the failure of case deletion approach in the deployment of prediction models.

One may consider that removing one or two sensors that have the largest amount of missing values from the dataset could solve the problem. That could help if mostly the same sensors were missing all the time. We can analyze in which way individual sensors are missing by the following experiment. First, we remove a sensor with the most missing values from the dataset, this way the observation vectors at each 30 min time stamp become shorter, they now include 35 sensors instead of 36. Given the updated observation vectors we recalculate how many of those vectors contain at least one missing value. Then we remove the next most missing sensor and repeat the calculation. Figure 2b presents the results. We see that removing a couple of largely missing sensors does not make the remaining observations complete. We would need to remove about half of the sensors in order to reach the stage where at least 95 % of the data are complete. The problem with this approach is that sensors to be removed may carry important information about the target, which would be lost if a lot of those sensors are removed. Figure 2c presents relation between the missing data rate in each sensor and the information about the target contained in it, measured as the absolute linear correlation with the target variable. We have removed the periods where the value of the target is equal to zero (the dark periods when there is no solar radiation)

AMTD
7, 7137–7174, 2014

from this analysis. We see some sensors in the far right corner and upper center that have high missing value rate, but also high correlation with the target variable. This means that excluding sensors with high missing value rates would lead to losses of valuable information about the target that would be useful for nowcasting.

One more issue with the data is that sensors produce missing values not independently from each other. For example, if one temperature value is missing, then it is likely that the other temperature values are missing as well. It may be the case that sensors are missing together due to some common external reasons, for instance, electric power outages. This observation is illustrated by Fig. 3, which plots pair-wise correlations between missing values for different sensors. Sensors that often are missing together are encoded in black (dark). We see that particularly temperatures (T), relative humidity (RH), visibility and precipitation readings are often missing together. This means that we cannot rely on redundancy of the sensors such that if say a temperature reading is missing at 33 m, we can use the reading at 50 m. Both readings would often be missing together.

Finally, in many cases the average duration of missing values lasts for several hours. Figure 4 presents the average duration of missing values in the case study dataset for each sensor. Since values may be missing for extended periods of times, from this perspective we also cannot simply discard data with missing values, since in such cases we often would not have model outputs for extended periods of time.

In summary, the amount of missing data is very large, at this level data with missing sensors cannot be discarded without losing valuable information. Missing values are strongly correlated with each other that makes it difficult and in many cases impossible to make use of sensor redundancy or impute missing data based on non-missing data. Removing sensors with the most missing data is also not feasible, since missing values are not concentrated in several sensors, but they are distributed across all the sensors and the sensors with a lot of missing values at the same time carry relatively strong information about the target at times when the values are not missing. Hence, the most appropriate solution to the problem of missing values in this setting appears to

AMTD
7, 7137–7174, 2014

be building models that are robust to missing data. This approach is free from any assumptions about the missing data and allows nowcasting even when all or nearly all the sensors are missing.

3.2 Prediction accuracy

Next we experimentally analyze accuracies of several linear regression models and their robustness to missing values. The first experiment demonstrates how we can select the best model for deployment. The second experiment presents evidence about the performance on unseen data.

Table 3 presents the errors of the regression models ALL, rALL, SEL, rSEL, PCA, rPCA, PLS measured on the training set using 5-fold cross validation and deterioration index estimated on the training set. For PCA, rPCA and PLS the number of components was fixed to $k = 18$, which is a half of the original number of input sensors, and explains 99 % of the variance. The cumulative percent variance method was used for selecting k, which is recommended as one of the most reliable methods in the literature (Valle et al., 1999). Figure A1 in the Appendix provides a complementary information about the variance explained by PCA components. Later in this section we will provide a sensitivity analysis to different values of k.

This analysis is performed from the perspective on an analyst, making a decision on which model to deploy. Cross validation is used to avoid potential overfitting of the model parameters to the training data. Complementary information on the goodness of fit and confidence intervals of the regression coefficients is presented in Appendix B.

In case the analyst bases the decision only on the offline analysis of validation errors, she would select ALL for deployment, since it shows the lowest error, while PCA and rPCA show nearly the highest error. However, the deterioration indexes computed for these models suggest the opposite: rPCA shows the best, while ALL shows the worst deterioration index value. The analyst now can theoretically compare the robustness of

AMTD
7, 7137–7174, 2014

two models, for instance, ALL and PCA, using the criteria from Eq. (11), which gives

$$m^\star = (r-1)\frac{[\mathrm{RMSE}^{(\mathrm{PCA})}]^2 - [\mathrm{RMSE}^{(\mathrm{ALL})}]^2}{d^{(\mathrm{ALL})} - d^{(\mathrm{PCA})}}$$
$$= (36-1)\frac{(21.8)^2 - (19.0)^2}{1\,122\,710 - (-117)} \approx 10^{-4}.$$

The result $m^\star = 10^{-4}$ means that, if we expect at least one sensor reading to be missing in ten thousand observations, it is better to deploy rPCA than ALL. Recall that in the data about 2.4 sensors are missing on average in every observation. Hence, in this situation it is clearly worth deploying rPCA, instead of a standard linear regression, even though the ordinary regression may be more accurate when no data is missing.

Let us consider the regularized version of ordinary regression rALL and rPCA. rALL shows better cross-validation accuracy than rPCA on the training data, and not that bad deterioration index, as ALL. For rALL and rPCA $m^\star = 0.4$, which means that it is still worth deploying rPCA.

How about non regularized PCA? The performance of PCA and rPCA seems very close to each other. For PCA and rPCA $m^\star = 13.1$, which means that rPCA is expected to be more accurate if more than 13 sensors are missing, which is a bit too pessimistic for our case study data. Hence, the analysis suggests to chose PCA for deployment. the test set, which we analyze.

The following analysis simulates online operation after deployment. The regression models are trained on the training set and then sequentially tested on the test set. Table 4 reports the testing results of the regression models ALL, rALL, SEL, rSEL, PCA, rPCA, PLS ($k = 18$).

The regularized principal component regression rPCA demonstrates the best performance on the test data (RMSE = 19.49), closely followed by PCA without regularization (RMSE = 19.52). The other regularized approaches outperform rSEL and rALL perform notably worse (RMSE = 20.43 and 20.28), but still outperform the naive baseline NAI (RMSE = 22.88). The unregularized approaches PLS, SEL and ALL perform

AMTD
7, 7137–7174, 2014

much worse than the baseline and illustrate well the dangers presented by massively missing values.

It is interesting to note that the analyzed strategy combining linear regression with mean replacement Žliobaitė and Hollmén (2013) theoretically approaches NAI performance as more values go missing. If all the input values are missing, then the predictor becomes NAI automatically.

To analyze the performance further we divide the test data into non-missing (44 %) and missing observations (56 %) parts and inspect the errors on these sub-sets separately. We see that the performance data of all the models is similar when there is no missing data, with the ordinary regression ALL having an advantage in accuracy, since it does not discard any information from the input data. However, the non-regularized models (ALL, SEL and PLS) fail badly when there is missing data, while the regularized rSEL and rALL lose some accuracy, but still remain competitive. Both non-regularized and regularized PCA remain nearly insensitive to missing data.

Figure 5 plots the distribution of absolute residuals for each approach. We can see that most of the errors (residuals) are concentrated around 10, which is not bad, given that the range of the target variable is from 0 to 100. It means that most of the predictions do not deviate too much from the true. We can also see that NAI has less probability mass on the left hand side, where the most accurate predictions are. As expected, intelligent predictors do better than NAI. Only the unregularized approaches ALL and SEL have any probability mass on the far right, which means that they occasionally produce predictions that may exceed the maximum of the true target. We can conclude from this investigation, that predictions by most of the approaches are reasonably stable, and outliers in predictions do not pose any major threats.

Next, let us analyze sensitivity to the parameter setting. So far we used a fixed number of components ($k = 18$) for PCA, rPCA, PLS and the same number of selected features for SEL and rSEL. Figure 6 the testing errors (RMSE) as a function of k.

An important observation can be made from this plot. The regularized approaches rSEL, rPCA performs reasonably well at all variants of the parameter k, while the

AMTD
7, 7137–7174, 2014

non-regularized models SEL, PCA and PLS perform poorly when a large number of components is retained. In such a case the resulting models are still similar to ALL, which uses all the available information. ALL, rALL and NAI do not depend on the parameter k, but are also included for comparison. We also observe that PLS becomes very effective at low k, but there is a risk of setting k incorrectly (e.g. around 25) in which case PLS gives the worst results. Therefore, we rather recommend using rPCA, which gives stable and accurate results even in k is sub-optimal.

Finally, we visually analyze the model outputs made by the baseline approach ALL and the recommended regularized approach rPCA ($k = 18$). Figure 7 plots four 3 day snapshots from year 2012: 1–3 January, 1–3 April, 1–3 July and 1–3 October. It is important to emphasize that here we plot the raw outputs of the classifiers to better illustrate the effects of regularization, whereas when calculating the numerical errors we post-process all the model outputs to fall into the same interval as the original target ([0, 100], where 0 means no irradiance is observed, and 100 (%) means all the theoretically possible irradiance is observed). That is, if the prediction is less than 0, we correct it to 0, and if the prediction is larger than 100, we correct it to 100. That makes the baseline classifiers more competitive (and hence is more prudent way of quantitative evaluation). We see from the figure that the baseline ALL sometimes fails to extremes (particularly in January and April plots), while the regularized approach rPCA remains stable. In July there are only a couple situations when ALL has very poor performance (we see a green inclination on day #2 and a green peak on day #3). In October both approaches perform similarly. Unlike ALL, rPCA perform stable and does not show extreme failures.

4 Summary and conclusions

In environmental monitoring, continuous and comprehensive measurement of the environment leads to streaming data setting. Nowcasting in such settings is a demanding task. We performed a case study in modeling solar radiation based on SMEAR

AMTD
7, 7137–7174, 2014

measurement data set, where model outputs are expected to be available continuously in spite of often missing sensor readings. We also experimentally analyzed the missing data patterns in our data set.

We aimed at nowcasting the amount of global radiation, relative to the theoretical maximum, with the help of measured meteorological variables. Due to the need to provide instantaneous outputs in the data streaming setting, as well as limited computing power, especially when operating on autonomous power sources, we dismiss any of the sophisticated data imputation methods, which are computationally more demanding. We experimentally analyzed accuracies and robustness to missing data of seven linear regression models, and recommend using regularized PCA regression. The results apply to linear regression models coupled with replacement of missing values by a constant (mean).

The strategy that we consider does not require any sophisticated missing value imputation, just replacing the values with pre-defined constants. Linear regression is also very light computationally, it only requires r multiplications, where r is the number of input variables, and one summation. When the model is trained, it can be recorded and operate with minimal energy consumption. If, in addition to that, we consider a computationally heavy imputation procedure, such as Expectation Maximization algorithm, it would require by orders of magnitude more computing power, and would be the dominating computing operation.

The regression itself is a powerful model, particularly considering that, if desired, one could apply non-linear transformations to the input features, which then would make the resulting predictions non-linear with respect to the inputs. More importantly, linear models are theoretically well understood, and can provide guarantees with respect to performance when there is a lot of missing data. We would argue that in such situations robustness of the model may be more important than flexibility. A flexible model may on average be more accurate, but the outputs may be extremely wrong at times. On the other hand, a robust model may be not the most accurate on average, but its performance would be stable and the errors not too large at all times. We chose linear

AMTD
7, 7137–7174, 2014

models, since they have theoretical guarantees for robustness. Hence, we recommend using our established guideline in training regression models, which themselves are robust to missing data.

Appendix A: Parameter selection

Figure A1 presents information on the variance explained by PCA components.

Appendix B: Goodness of fit

Table B1 presents fitness statistics of the regression models to the training data. The coefficient of determination R^2 indicates the amount of total variability explained by the regression model. The coefficient is computed as

$$R^2 = 1 - \frac{\sum_{l=1}^{n} (\hat{y}^{(l)} - y^{(l)})^2}{\sum_{l=1}^{n} (y^{(l)} - \overline{y})^2},$$

where $y^{(l)}$ is the true target value of the lth sample and $\hat{y}^{(l)}$ is the corresponding model output, $\overline{y}$ is the mean of the true target values, and n is the number of samples in the train set. We see that the best fit model is ALL. Recalling the experimental analysis in Sect. 3 we can see that good fitness to the training data does not guarantee good generalization performance when a lot of missing values start to appear.

The Supplement related to this article is available online at doi:10.5194/amtd-7-7137-2014-supplement.

Acknowledgements. This work has been supported by the Academy of Finland grant 118653 (ALGODAN), and grant 258568 (MultiTree).

AMTD
7, 7137–7174, 2014

References

Allison, P.: Missing Data, Sage Publications, 2001. 7139

Bacher, P., Madsen, H., and Nielsen, H. A.: Online short-term solar power forecasting, Sol. Energy, 83, 1772–1783, 2009. 7140

Bhardwaj, S., Sharma, V., Srivastava, S., Sastry, O., Bandyopadhyay, B., Chandel, S., and Gupta, J.: Estimation of solar radiation using a combination of Hidden Markov model and generalized Fuzzy model, Sol. Energy, 93, 43–54, 2013. 7140

Black, C., Broadstock, D., Colin, A., and Hunt, L. C.: Filling in the gaps in transport studies: a practical guide to developments in data imputation methods, Traffic Eng. Control, 48, 358–363, 2007. 7143

Enders, C. K.: Applied Missing Data Analysis, Guilford Press, 2010. 7143

Hammer, A., Heinemann, D., Lorenz, E., and Lückehe, B.: Short-term forecasting of solar radiation: a statistical approach using satellite data, Sol. Energy, 67, 139–150, 1999. 7140

Hari, P. and Kulmala, M.: Station for Measuring Ecosystem-Atmosphere Relations (SMEAR II), Boreal Environ. Res., 10, 315–322, 2005. 7138, 7140

Hastie, T., Tibshirani, R., and Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001. 7144, 7145, 7146

Hoerl, A. E. and Kennard, R. W.: Ridge regression: biased estimation for nonorthogonal problems, Technometrics, 12, 55–67, 1970. 7146

Hrust, L., Klaic, Z. B., Krizana, J., Antonic, O., and Hercog, P.: Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations, Atmos. Environ., 43, 5588–5596, 2009. 7139

Jolliffe, I. T.: Principal Component Analysis, 2nd edn., Springer, 2002. 7146

Junninen, H., Niska, H., Tuppurainen, K., Ruuskanen, J., and Kolehmainen, M.: Methods for imputation of missing values in air quality data sets, Atmos. Environ., 38, 2895–2907, 2004. 7139, 7144

Junninen, H., Lauri, A., Keronen, P., Aalto, P., Hiltunen, V., Hari, P., and Kulmala, M.: Smart-SMEAR: on-line data exploration and visualization tool for SMEAR stations, Boreal Environ. Res., 14, 447–457, 2009. 7141

Kadlec, P., Gabrys, B., and Strandt, S.: Data-driven soft sensors in the process industry, Comput. Chem. Eng., 33, 795–814, 2009. 7139, 7143

AMTD
7, 7137–7174, 2014

Kopp, G. and Lean, J. L.: A new, lower value of total solar irradiance: Evidence and climate significance, Geophys. Res. Lett., 38, L01706, doi:10.1029/2010GL045777, 2011. 7142
Lu, H., Hsieh, J., and Chang, T.: Prediction of daily maximum ozone concentrations from meteorological conditions using a two-stage neural network, Atmos. Res., 81, 124–139, 2006. 7139
Marquez, R. and Coimbra, C. F.: Forecasting of global and direct solar irradiance using stochastic learning methods, ground experiments and the NWS database, Sol. Energy, 85, 746–756, 2011. 7140
Menut, L. and Bessagnet, B.: Atmospheric composition forecasting in Europe, Ann. Geophys., 28, 61–74, doi:10.5194/angeo-28-61-2010, 2010. 7139
Michalsky, J.: The Astronomical Almanac's algorithm for approximate solar position (1950–2050), Sol. Energy, 40, 227–235, 1988. 7142
Valle, S., Li, W., and Qin, S. J.: Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods, Ind. Eng. Chem. Res., 38, 4389–4401, 1999. 7153
Vuilleumier, L., Calpini, B., Cattin, R., Roulet, Y.-A., Stauch, V., Stöckli, R., and Giunta, I.: Solar radiation now-casting: the need for multiple data source integration, in: Proc. of the COST Action ES1002 "WIRE" State of the Art Workshop, available at: http://www.wire1002.ch/fileadmin/user_upload/Major_events/WS_Nice_2011/Spec._presentations/Vuillemier.pdf (last access: 14 July 2014), 2011. 7140
Wold, S., Sjostroma, M., and Eriksson, L.: PLS-regression: a basic tool of chemometrics, Chemometr. Intell. Lab., 58, 109–130, 2001. 7147
Žliobaitė, I., and Hollmén, J.: Fault tolerant regression for sensor data, in: Proc. of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECMLPKDD, 449–464, 2013. 7140, 7147, 7148, 7155
Žliobaitė, I., and Hollmén, J.: Optimizing regression models for data streams with missing values, Mach. Learn., 1–27, doi:10.1007/s10994-014-5450-3, 2014. 7140

AMTD
7, 7137–7174, 2014

Table 1. Sensors for the case study: SWS – surface wetness sensor, *P* – pressure, *T* – temperature, WS – wind speed, WD – wind direction, RH – relative humidity, RH Td -relative humidity calculated using dew point, PTG – potential temperature gradient, Vis – visibility.

index	measurement	height	missing values
1	Rain	18.0 m	1 %
2	SWS	18.0 m	1 %
3	Dew point	18.0 m	18 %
4	*P*	0.0 m	2 %
5	*T*	4.2 m	16 %
6	*T*	8.4 m	3 %
7	*T*	16.8 m	2 %
8	*T*	33.6 m	2 %
9	*T*	50.4 m	2 %
10	*T*	67.2 m	2 %
11	WS	33.6 m	9 %
12	WS	8.4 m	5 %
13	WS	16.8 m	3 %
14	WS	33.6 m	9 %
15	WS	74.0 m	25 %
16	WD avr		2 %
17	WD ultrasonic	8.4 m	7 %
18	WD ultrasonic	16.8 m	4 %
19	WD ultrasonic	33.6 m	9 %
20	WD ultrasonic	74.0 m	23 %
21	RH	4.2 m	21 %
22	RH	8.4 m	9 %
23	RH	16.8 m	7 %
24	RH	33.6 m	7 %
25	RH	50.4 m	9 %
26	RH	67.2 m	6 %
27	RH Td	18.0 m	20 %
28	PTG		5 %
29	Visibility	18.0 m	1 %
30	Vis-min	18.0 m	1 %
31	Vis-max	18.0 m	1 %
32	Precipitation intensity	18.0 m	1 %
33	Preci-min	18.0 m	1 %
34	Preci-max	18.0 m	1 %
35	Precipitation	18.0 m	1 %
36	Snowfall	18.0 m	1 %
37	Global RADIATION	18.0 m	1 %

AMTD
7, 7137–7174, 2014

Table 2. Summary of regression models: OLS – ordinary least squares, RR – Ridge regression.

		Optimization	
		OLS	RR
Inputs	all r	ALL	rALL
	selected k	SEL	rSEL
	PCA k	PCA	rPCA
	PLS k	PLS	

AMTD
7, 7137–7174, 2014

Table 3. 10-fold cross validation errors (RMSE) measured on the *training* dataset and deterioration index (d).

	ALL	rALL	SEL	rSEL	PCA	rPCA	PLS
RMSE	19.0	21.6	20.5	21.9	21.7	21.8	20.8
d	1 122 710	537	362 708	451	−109	−117	6121

AMTD
7, 7137–7174, 2014

Table 4. Nowcasting errors (RMSE) on the testing dataset.

	ALL	rALL	SEL	rSEL	PCA	rPCA	PLS	NAI
full set	175.8	20.4	127.8	20.3	19.5	19.5	25.7	22.9
non-missing	17.9	19.2	18.8	19.7	19.6	19.6	19.3	22.9
missing	233.4	21.3	169.2	20.7	19.4	19.4	29.7	22.9

AMTD
7, 7137–7174, 2014

Table B1. Fitness statistics of the models on the training data.

	ALL	rALL	SEL	rSEL	PCA	rPCA	PLS
RMSE	19.2	21.2	20.5	21.6	21.8	21.8	20.9
R^2	0.501	0.393	0.436	0.373	0.361	0.361	0.410

AMTD
7, 7137–7174, 2014

Algorithm 1: PLS regression

Data: training set $(\mathbf{X}, \mathbf{y})$, number of components k
Result: estimated regression coefficients $\hat{\boldsymbol{\beta}}_{PLS}$

for $i \leftarrow 1$ **to** k **do**
  $\mathbf{w}_i \leftarrow \mathbf{X}^T\mathbf{y} / \sqrt{\mathbf{y}^T\mathbf{X}\mathbf{X}^T\mathbf{y}}$;
  $\mathbf{t}_i \leftarrow \mathbf{X}\mathbf{w}_i$;
  $q_i \leftarrow \mathbf{t}_i^T\mathbf{y} / (\mathbf{t}_i^T\mathbf{t}_i)$;
  $\mathbf{p}_i \leftarrow \mathbf{X}^T\mathbf{t}_i / (\mathbf{t}_i^T\mathbf{t}_i)$;
  $\mathbf{X} \leftarrow \mathbf{X} - \mathbf{t}_i\mathbf{p}_i^T$ (data deflation step);
  $\mathbf{y} \leftarrow \mathbf{y} - \mathbf{t}_i q_i$ (data deflation step);
end
$\mathbf{W} \leftarrow (\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_k)$;
$\mathbf{P} \leftarrow (\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_k)$;
$\mathbf{q} \leftarrow (q_1, q_2, \ldots, q_k)^T$;
$\hat{\boldsymbol{\beta}}_{PLS} = \mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1}\mathbf{q}$

Algorithm 1. PLS regression.

AMTD
7, 7137–7174, 2014

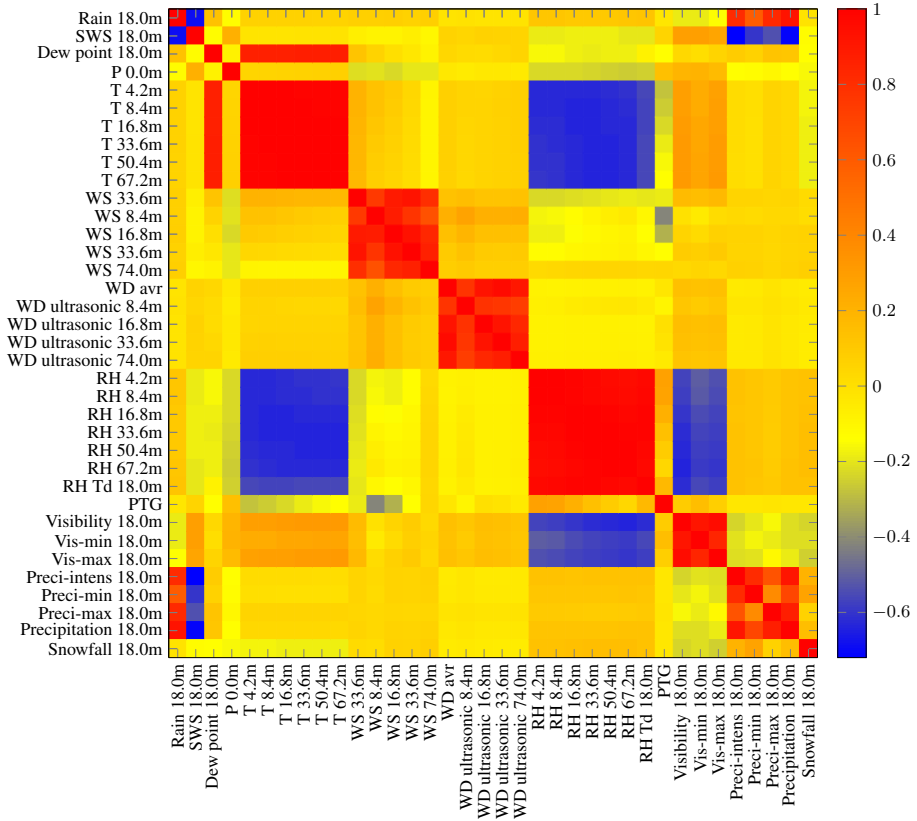

Figure 1. Correlations between input sensors.

AMTD
7, 7137–7174, 2014

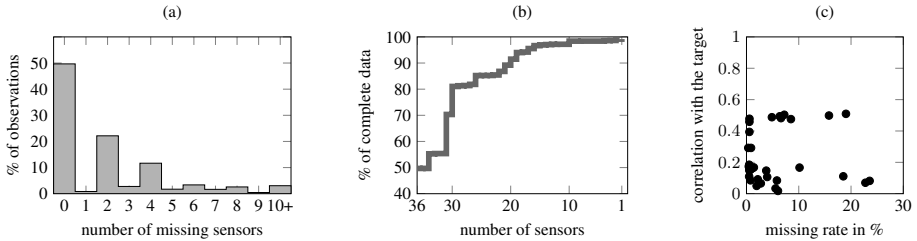

Figure 2. Analysis of missing data patterns: **(a)** distribution of number of missing sensors in observations (10+ means from 10 to 36 are missing); **(b)** effects of removing the most missing sensors, **(c)** the relation of individual sensors with the target variable (each dot represents one sensor).

AMTD
7, 7137–7174, 2014

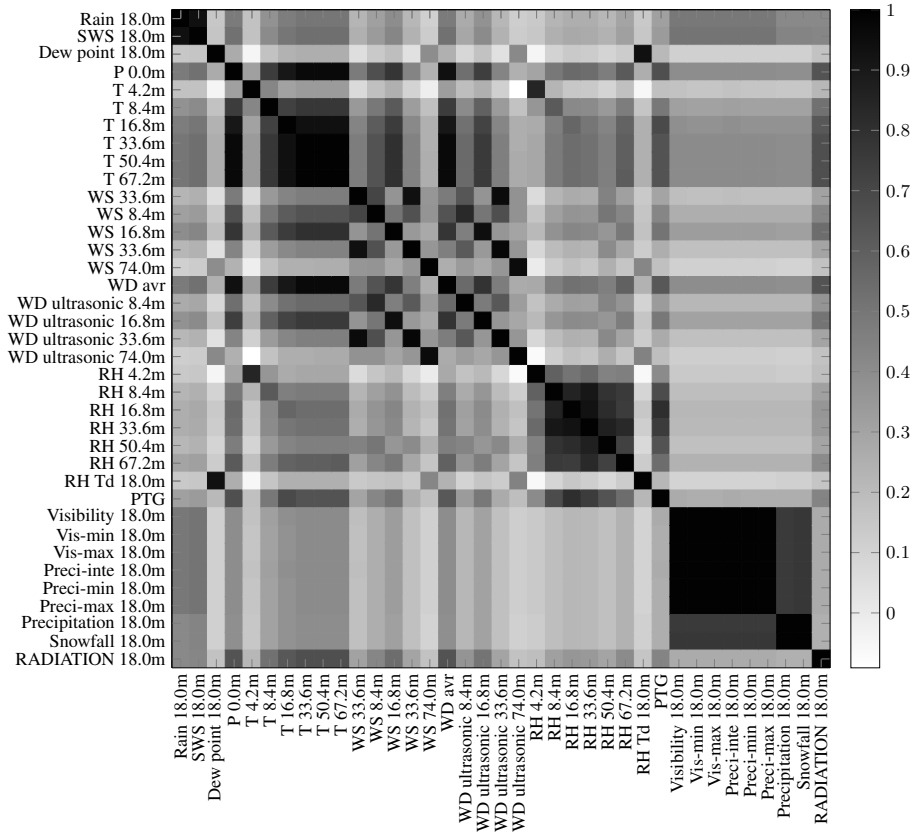

Figure 3. Correlation of missing value patterns. High correlation (indicated by darker values) means that the values are often missing together.

AMTD
7, 7137–7174, 2014

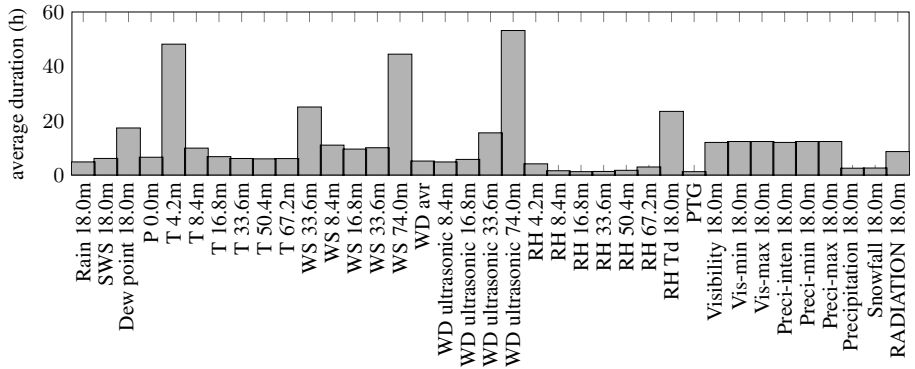

Figure 4. Average duration of missing readings.

AMTD
7, 7137–7174, 2014

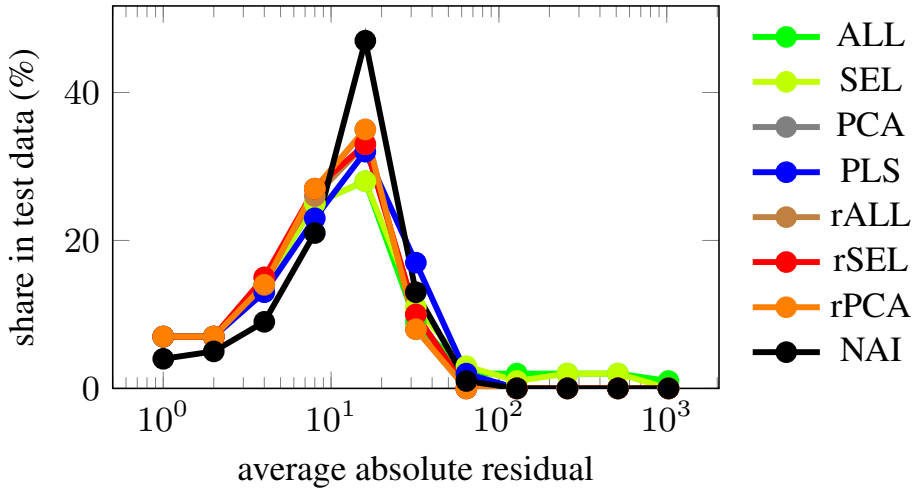

Figure 5. Analysis of residuals.

AMTD
7, 7137–7174, 2014

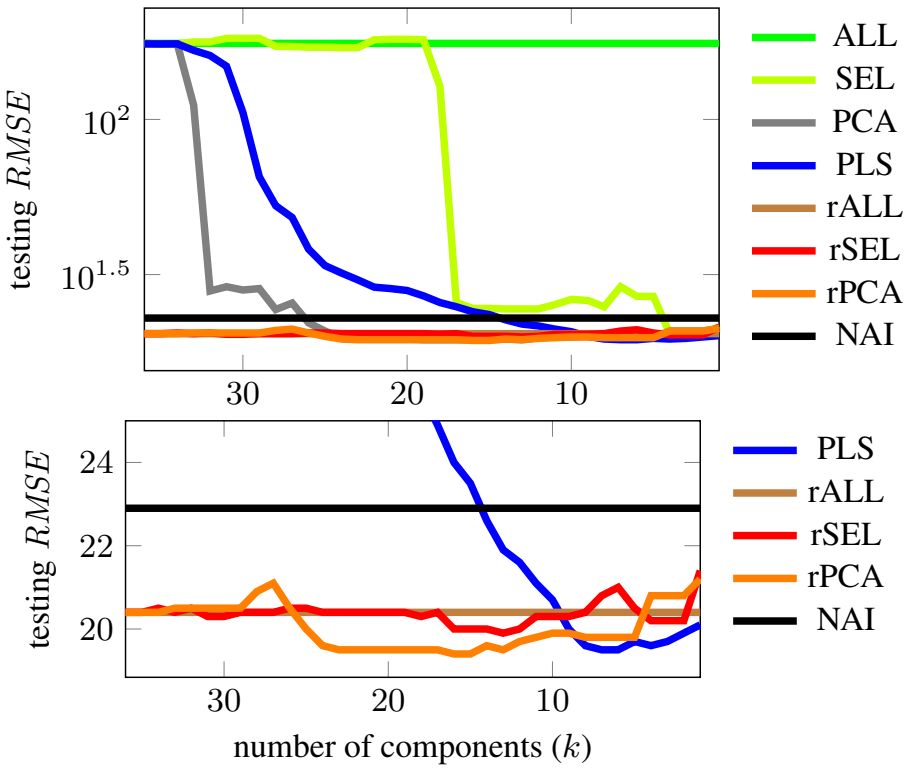

Figure 6. Nowcasting error as a function of components retained (top plot – all models in log scale, bottom plot – best models zoomed in).

AMTD
7, 7137–7174, 2014

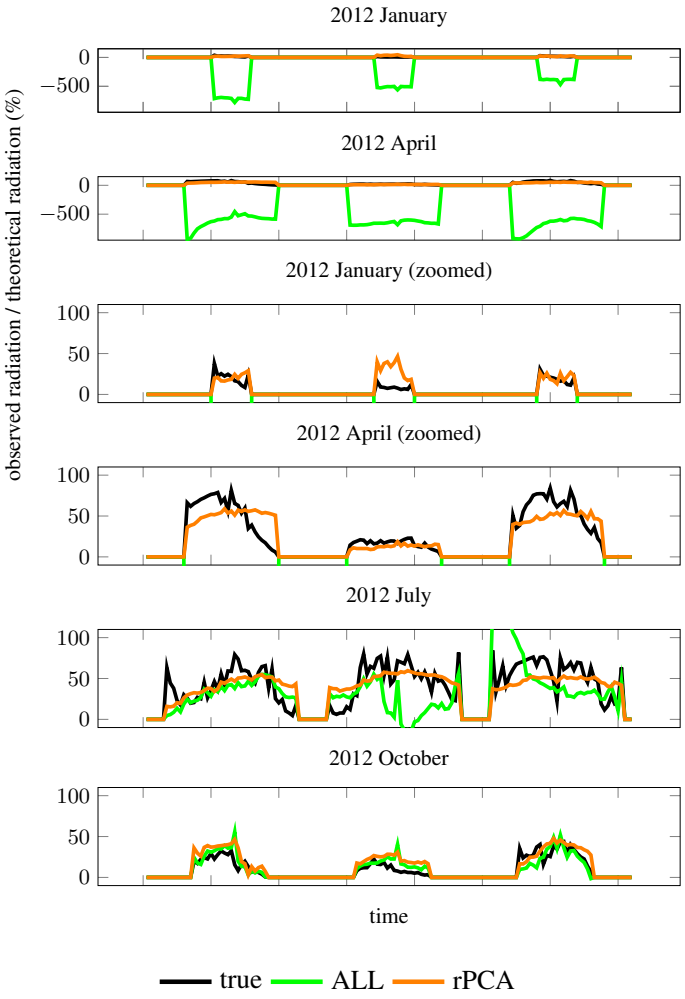

Figure 7. Visualization of nowcasting (each plot shows 3 days).

AMTD
7, 7137–7174, 2014

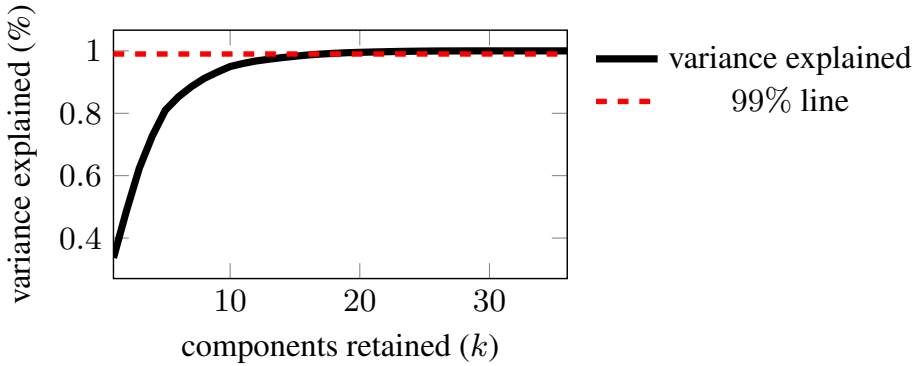

Figure A1. Cumulative variance explained by PCA components on the training data without missing values.

AMTD
7, 7137–7174, 2014