Network Management System Dimensioning with Performance Data

Kaisa Tuisku

University of Tampere
School of Information Sciences
Computer Science
M. Sc. thesis
Supervisor: Jorma Laurikkala
June 2016

University of Tampere
School of Information Sciences
Computer Science
Kaisa Tuisku: Network Management System Dimensioning with Performance Data
M. Sc. thesis, 59 pages, 9 appendix pages
June 2016

The goal of this work was to develop a new solution for dimensioning Nokia's network management system by utilizing performance data to achieve more accurate predictions for capacity usage. An algorithm for generating predictive models for different system resources and application performance attributes was implemented. Multiple linear regression with an exhaustive branch-and-bound search through the space of possible predictor variables was utilized to predict the usage of different resources. Solutions for the internal and external validation, outlier detection, and the visualization of the models were also specified. The algorithm was evaluated with a sample of data from real network management system environment and the results were promising, because fairly accurate predictions could be made. This solution is still in progress: The model maintenance and specification, as well as handling nonlinear relationships must be considered.

Keywords: network management system, dimensioning, performance, data analysis, multiple linear regression, subset selection.

Contents

1. Introduction

Data mining techniques (Hand *et al.*, 2001) have lately broadened the range of techniques for data analysis. The goal in data mining is to find unexpected patterns between variables or to summarize the big data sets with simple and understandable models. Since the amount of available data has recently grown tremendously, new information is now available for us to find and use for our own benefit in numerous areas. The amount of data has grown also in network management performance field, due to new generations of mobile telecommunications technologies which have brought more and more variables to the system (Rémy & Letamendia, 2014). Using data in performance analysis and dimensioning purposes has become not only possible, but also necessary in network management system framework.

A network management system is an application for managing and monitoring network components, satisfying at least some of the systems management requirements listed in management framework system standards (International Organization for Standardization, 1989). Dimensioning a network management system means defining the minimum capacity that still allows the requirements for the system to be fulfilled. Dimensioning is used for determining which network management system configuration or scaling in cloud should be chosen to the network managed by the customer. The concept of dimensioning is closely related to the concept of system performance. The system performance consists of hardware dimensioning parameters and workload parameters which determine the system capacity.

In this thesis, a data-driven solution for network management system dimensioning is introduced. Since this approach is new to the product, the solution with all the required functionality is not in the scope of this thesis. However, the core algorithm with general guidelines to continuation is implemented. The main idea was to utilize the performance data of the network management system for predicting capacity usage and to use the information for more flexible and accurate dimensioning. However, for achieving trustworthy predictions, the important issue of uncertainty when mapping the seldom perfect or completely predictable real world to data matrixes must be recognized (Hand *et al.*, 2001). The final goal is to develop a new algorithm for dimensioning modeling to solve problems risen with the previous dimensioning solutions and to answer future needs, while still understanding the limitations of data-driven solutions.

Historically, many different dimensioning solutions for network management systems exist. In early the 1990's, after the launch of the second

generation of wireless telephone technology, the first version of Nokia's network management system was released. The earliest releases had one configuration, in which customers were instructed not to cross predetermined limits for capacity usage. Later on in the 1990's, more configurations were offered. The numbers of second generation transceivers, mobile switching centres and home location register elements determined the capacity of each configuration. Even though both the networks and network management systems were a lot simpler than nowadays, the performance of the system was poorly understood, because early state network simulators became available only mid-nineties (Kuusela, 2015).

The 21th century was a turning point in network technology. The functionality of the earlier network management systems had to be expanded with the support for third generation mobile network functionality, such as mobile internet access. Few configuration options were still offered, but it became necessary to increase the amount of hardware resources, such as memory and central processing units, inside a configuration. This created a need for more accurate dimensioning.

A programmable solenoid controller simulator was used with data from real network elements to perform tests with different amounts of hardware resources. After some data points from simulator tests were discovered, the empty spaces in between were filled with values from a straight linear line which was drawn based on the few measured data points. According to T. Kuusela (personal communication, August 12, 2015) it soon became clear that estimating the configuration and the increase in performance was very time consuming. Therefore, the first multiple linear regression based dimensioning tool, with manually estimated coefficients, was created.

By 2010, the fourth generation mobile network technology brought new functionality, such as mobile broadband internet access and voice services. The number of different types of network elements increased tremendously. At the same time, big changes to the capacity statement were implemented. The system became so complex that understanding the structure was very hard, being still simpler than nowadays. The customers also demanded more accurate information of the product performance than earlier. At start, capacity estimation was done with the same method as before, but as the new network management system was released in 2011, performance data were utilized instead of manual estimation in multiple linear regression modeling. This provided more precise estimations, but showed that more performance data from varying environments were needed. This proved to be a challenging task (Heinonen, 2015a).

The current version of the network management system and its successor releases brought completely renewed system architecture in 2014. Since the network management system and the managed networks again became more complex, the capacity estimation method of the sixth version did not work with the next one. At first, linear regression modeling was again attempted, but it soon became too complicated to implement due the lack of data, non-optimal predictor variable choices, and unsuitable tools. Thus, according to Heinonen (2015a), a new simplified system was quickly developed to be able to offer at least the same functionality as before.

Due to the more complex environment and lack of time in the implementation phase, the current dimensioning tool has some consequential shortages (Heinonen, 2015a). Firstly, the modeling of both the capacity and the produced load is somewhat inaccurate, due to extreme simplification. Secondly, the publishing environment, a spreadsheet application, of the current tool is inflexible. Because of these issues each new product release required new test data to build new models on. Therefore, a more versatile and accurate dimensioning tool demanding less maintenance is necessary.

Future challenges must also be paid attention to in developing a new dimensioning solution. First of all, cloud environment in which the products are likely to move on in near future allows more customer-specific configurations. To make use of this possibility, the capacity usage must be better known. Secondly, the more and more complex networks the customers are working with may be economically demanding. Therefore, the efficient usage of hardware must be taken in consideration. The dimensioning tool currently in use is not accurate enough to answer to these needs.

The new solution for network management system dimensioning is designed to answer the challenges set by the shortages of the current and needs of the future dimensioning tool. For predicting the capacity usage, multiple linear regression method was chosen again, since the method has been working well in this environment once the choices for predictor variables are correct. Multiple linear regression method is also highly extensible to nonlinear relationships, if those are detected to describe a resource better than a linear model. The simplicity of the multiple linear regression model is also an advantage, because besides predicting the capacity usage, understanding the system dependencies is a useful, even necessary, side effect of the modeling algorithm. With these advantages, the multiple linear regression model was chosen over, for instance, the more complicated multivariate linear regression model and the less understandable neural network model.

Since one requirement for the solution was that it should be data-driven, the selection of the prediction variables to be included in a model must be based on the available data instead of predetermination. After exploring some subset selection methods, it was finally decided that only exhaustive methods give results that are trustable enough. As exhaustive calculation is computationally demanding, the branch-and-bound method, where some branches can be excluded from calculation, was preferred over the other exhaustive subset selection methods.

Besides selecting the analysis methods, work involving the data preparation and the final algorithm had to be done. These include initial data analysis to ensure a good fit of the chosen model. Also, the pre-processing of the data was a time-consuming task, because much had to be learned and specified before any type of automation could be implemented. Selecting a suitable tool to implement the final algorithm with all the necessary features was also an important part of the work. In this thesis, the whole data mining process (Hand *et al.*, 2001) from the initial steps to the implementation of the dimensioning tool and the evaluation of final outcome was carried out.

Based on a test run with data from the usage of one resource of the network management system, the algorithm seems to work quite well. Linear relationships between the response variable and the predictor variables were detected, and the best of the models could explain approximately 75 % of the variation in data. Still, some issues were detected in test run, but it is very likely that the issues are solved once more data with more variation are available. Similar evaluation will be carried out with every system resource and other performance attributes once the data are available. This may lead to finding new issues with the algorithm. However, at this point of the work the results look promising.

The theory behind the new network management system dimensioning solution is presented in Chapters 2–4. Chapter 2 is for the network related terminology and concepts. Chapter 3 presents the statistical model, multiple linear regression method for modeling resource usage and other performance indicators of interest. The problem of subset selection in regression analysis in reviewed in Chapter 4. Chapters 5 and 6 concentrate on what was done in practise: Chapter 5 gives an insight to the preparatory work and Chapter 6 describes the resulting algorithm. Discussion on the subject as well as a test run with sample data is presented in Chapter 7, while in Chapter 8 the possible future work is described. Finally, the outcome of the work is concluded in Chapter 9.

2. Network management

The model and framework from Telecommunications Management Network (International Organization for Standardization, 1989) is an international standard for network management having modules for fault, configuration, accounting, performance, and security (FCAPS). The network management system on which this work concentrates on is both for network and network element management and includes four of these five network management categories. Only accounting is left to the customer to carry out themselves. Every module category has a range of centrally accessed management functions, applications, and interfaces. All functionality is accessed centrally via a graphical user interface.

For the newest version of the network management system, there are three existing configurations to match the requirements of the different sized networks. The configurations differ from each other by the number of managed integrated network elements, allocated hardware resources, and used virtual infrastructure setups. This is related to the main concept of this thesis, dimensioning. The idea of dimensioning is presented in Section 2.3, but for deeper understanding, the architecture and functionality of the particular network management system are first described in Sections 2.1 and 2.2.

2.1. Architecture

In Nokia's earlier network management systems, all the functionality was paired with dedicated hardware, and services were allocated on physical servers (Nokia Solutions and Networks, 2015). From the first release of the current version, the system has been working on top of virtualized infrastructure, and physical servers act as a combined resource pool. The functionality is nowadays distributed to multiple virtual machines. In comparison to traditional infrastructure, the advantages of virtualization are clear: more efficient resource and hardware usage, reduced downtime, and more accurate scalability. On the other hand, defining the minimum capacity requirements is a substantially more complex process in the virtualized environment.

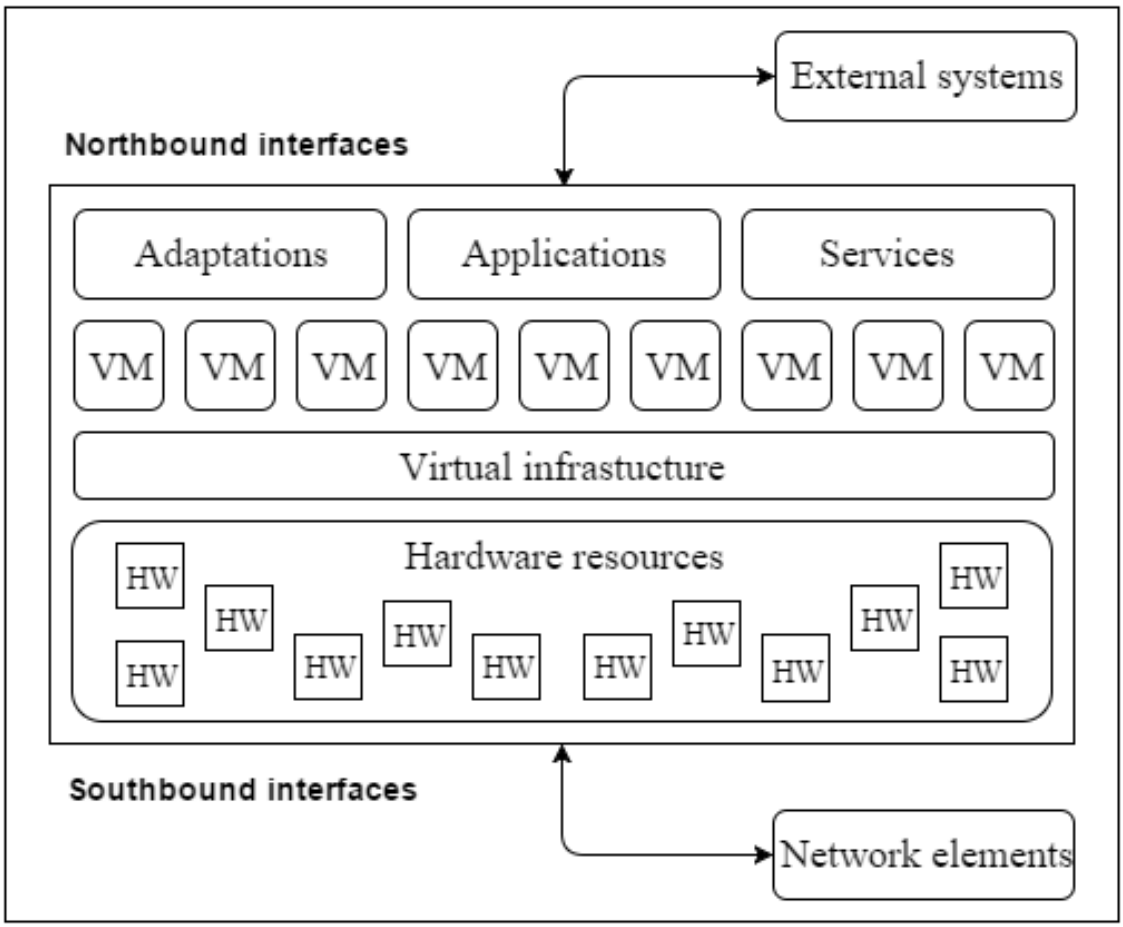

Figure 1. Nokia's network management system architecture.

As presented in Figure 1, the system is connected to network elements and lower-level systems, such as element manager systems or mediators, through southbound interfaces. The physical resources act as a resource pool for the whole infrastructure. The virtualized infrastructure is the link between hardware resources (HW) and virtual machines (VM): It creates and maintains the virtual machines on top of the physical hardware. The virtual machines are then accessed by adaptations, applications, and services. Finally, the northbound interfaces are used for accessing users' workstations, users' external systems, and third-party software securely within the network management system.

2.2. Modules

The fault, configuration, performance, and security modules of the network management system contain different tasks (Nokia Solutions and Networks, 2015). The three main tasks of the *fault management module* are to control the network monitoring process, to solve the most essential problems, and to detect and troubleshoot the disruption in the network services. The network monitoring system collects and processes alarms, the network fault indicators.

When a fault occurs somewhere in the network, an alarm is created. One network problem can cause multiple alarms in different network elements. The fault management system reduces the amount of the alarms by filtering, reclassifying, and compressing highly correlating alarms together, allowing the user to notice the real source of problems and start working on the solutions quicker. The network, or parts of it, can also be visualized based on geographical areas, workstation networks or transmission networks to trace alarms.

The *configuration management module* is for identifying characteristics, provisioning changes, and verifying compliance with specified requirements of the network configuration data. The characteristics of configuration data are identified by viewing the consistency reports of different configurations managed by the system. Changes are provisioned by managing and editing different network elements and configurations. Compliance verification supports the starting and scheduling of different operations, such as the exporting and importing of files and the uploading and downloading of data. Reference configurations are managed in configuration management module category.

The aim of the *performance management module* is to collect data for various activities. These include monitoring network functions and subscriber behavior, verifying the configuration of the telecommunications network, localizing potential problems, and providing services to mobile subscribers. Online-oriented performance monitoring displays real-time information on the network performance which is used mainly for accessing information of problematic cases. Offline-oriented performance reporting applications display information of the performance of the system over a chosen time period. This information can be used for troubleshooting, planning or optimizing the network.

The *security management* serves the guidelines issued on confidentiality, integrity, and availability. These guidelines are realized by system hardening, user security, network security, and security supervision. The goal of these functions is to enforce and manage the security related information, and to maintain the general security of the system.

2.3. Dimensioning

Dimensioning in its traditional context is, alongside detailed radio system planning and optimization, one of the main phases in radio system planning process (Lempiäinen & Manninen, 2002). The goal of dimensioning is to initially draft the radio network configuration and deployment strategy, and to

define the essential parameter values and technologies. The dimensioning of the network management system studied in this thesis is a similar process. As a result of the dimensioning phase, the minimum capacity that still allows the requirements for the system to be met is defined.

One target of Nokia's network management system dimensioning is to determine a suitable system configuration or scaling in cloud for the network managed by the customer, based on the network topology and workload information (Nokia Solutions and Networks, 2015). On the other hand, the target can be to understand how much free capacity the existing network management system still has. These targets lead to the main goal of dimensioning the Nokia's network management system, which is the ability to get fairly accurate estimations for the processing capability of the system, the system performance.

System performance is considered good, when the system can handle the assigned tasks effectively, in timely manner. For the customer, system performance is related to user experience and usability, but for developers the subject is more complex. The factors having an effect on system performance, the performance parameters, can be divided to two categories: hardware dimensioning parameters and workload parameters. In Nokia's network management system, there are dozens of performance parameters to be considered in dimensioning.

Hardware dimensioning parameters describe the amount and the usage of hardware resources, such as central processing units or memory. The hardware determines the number of different types of operations, for instance the number of simultaneous users, the server can handle. In virtualized network management systems, the virtualized architecture operates independently from underlying hardware, while every virtual machine has a designated amount of hardware resources.

Both in radio network and in network management system dimensioning, the system workload parameters describe the traffic in the system. The radio network traffic is mainly user actions and signaling between the network and user equipment. In network management systems, the traffic consists of events from administrative users, network elements, and external systems. An event can be, for instance, a network failure, a configuration change or a network performance change.

3. Multiple linear regression

Multiple linear regression (Freedman, 2009) is used for both descriptive and predictive modeling (Hand *et al.*, 2001). The model is constructed from a sample of data points, in which all variables are known. The simplest linear regression method for model fitting is ordinary least squares, where the sum of squared error is minimized for model creation. In descriptive analysis the features of existing data are modeled. For predictive analysis, the value of the response variable of a new case is predicted based on the predictor variables. In multiple linear regression, the response variable is represented as a linear combination of p predictor variables:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon.$$

In the equation, Y represents the response variable and X's are the predictor variables. The intercept (β_0) is the point where the regression plane, or in simple linear regression the regression line, meets the Y axis. The other β's are the coefficients for predictor variables, and ε stands for the error of the model. Multiple linear regression should not be confused with multivariate linear regression model (Fujikoshi *et al.*, 2010), in which more than one response variables are being predicted at once.

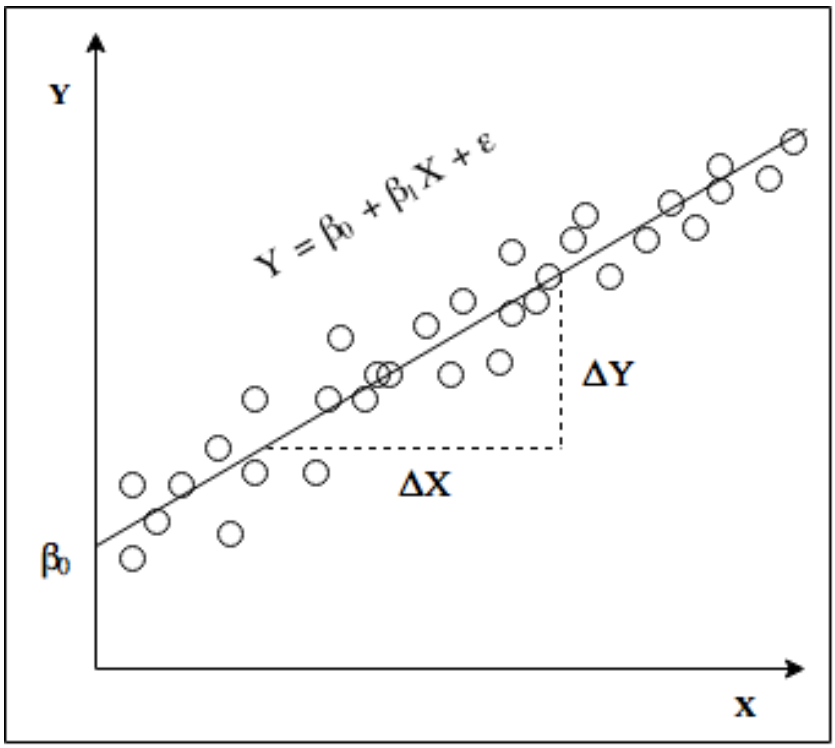

Figure 2. Regression line.

Figure 2 presents the *simple linear regression* function which has one response variable Y and one predictor variable X (p = 1). The function is easy to visualize in two-dimensional space: The Y axis represents the response variable and the X axis represents the predictor variable. The regression line is an illustration of the regression equation. The intercept of the equation is β_0 and the slope of the regression line, $\Delta Y/\Delta X$, is β_1, if the error ε is assumed to be zero. When the coeffiecient β_1 is zero, the Y value is the same as β_0. The variation orthogonal to the regression line is left unexplained by the model.

To make reliable predictions with multiple linear regression, where the number of predictor variables is two or more (p > 1), the data must fulfill more demands than in simple linear regression. First of all, the relationship between the response variable and the linear combination of the predictor variables has to be linear. The lack of linearity affects model fit indicators, errors and residuals, and weakens the results of statistical significance tests. Secondly, the predictors should not be correlated with each other, in order to prevent *multicollinearity* (Goldberger, 1991). Multicollinearity can be detected, for instance, by the variance inflation factor. Thirdly, some data points might be far away from others, often due to measurement errors or unmeasured variables. These data points are called *outliers* (Maddala, 1992). Hand *et al*. (2001) have also proven that with a small sample size, imprecise and biased estimates may appear. In this chapter some methods for ensuring reliable predictions are looked into: Measuring for the goodness of fit, handling multicollinearity, ensuring statistical significance, preventing erroneous modeling, and detecting the outliers of the model.

3.1. Statistical significance of model and predictors

In linear regression analysis, *t*- and *F-tests* (Kutner et al., 2005) for statistical significance are used for evaluating the model and its parameters. The results are for deciding whether the model is good enough for later use. When exploring the statistical significance of coefficient β_i, common hypotheses are

$$H_0: \beta_i = 0 \text{ and}$$
$$H_1: \beta_i \neq 0.$$

If the null hypothesis is rejected, there is a linear relationship between Y and X_i and the variable X_i may be a useful in the model. The t statistic is defined as

$$t = \frac{b_i}{s(b_i)},$$

where s(b_i) is the standard deviation of b_i , which is the point estimator of β_i:

$$b_i = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2},$$

in which $\overline{X}$ and $\overline{Y}$ are the means for X_i's and Y_i's. The test statistic follows the *t*-distribution with *n – p – 1 degrees of freedom* (*df*), when the hull hypothesis holds. The null hypothesis is either accepted or rejected based on the two-sided *t*-test statistics:

$$\text{If } |t| \leq t(1 - \alpha/2\ ;\ n - p - 1), \qquad \text{conclude } H_0.$$
$$\text{Otherwise,} \qquad \text{conclude } H_1.$$

In this formula, α is the significance level, that is, the probability to reject the null hypothesis, when it is true, n is the number of data points used for calculating the model and p is the number of variables in the model. A common value for α is 0.05, which means a five percent risk of concluding β_i to be significant, when it is not. Another method for choosing the correct hypothesis is by calculating the *p-value* and comparing it to the *α-level*. The *p*-value is the probability of observing a value as extreme as *t*. If the *p-value* is lower than α, the null hypothesis is rejected and the alternative one is accepted.

Evaluating the intercept of the model is slightly different from evaluating the coefficients. However, the alternative hypotheses appear similar, even though in this case a linear relationship is not evaluated. The goal is to find out, if the intercept could be zero:

$$H_0: \beta_0 = 0 \text{ and}$$
$$H_1: \beta_0 \neq 0.$$

The equation for the *t* statistics is the same as earlier:

$$t = \frac{b_0}{s(b_0)}.$$

However, the point estimator calculation for β_0 differs from that for β_i:

$$b_0 = \bar{Y} - b_1\bar{X}.$$

The conclusions are made similarly in the significance test for β_i, with the exception of degrees of freedom being $n - 2$.

The full model can be evaluated with the following *overall F-test*. The test detects whether there is a regression relation between the response variable and the linear combination of the predictor variables. The null and the alternative hypotheses are:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \text{ and}$$
$$H_1: \text{not all } \beta_i \text{ are equal to zero.}$$

The test statistic is

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)},$$

where *MSR* stands for the regression mean squared, while *MSE* is for the error mean squared. The SSR, the regression sum of squares, is defined as follows:

$$SSR = \sum (\hat{Y}_i - \overline{Y})^2,$$

where the $\hat{Y}_i$ is the predicted value and the $\overline{Y}$ is the mean. The *SSE*, on the other hand, stands for the error sum of squares:

$$SSE = \sum (Y_i - \hat{Y}_i)^2.$$

The n is the size of the sample and the p is the number of variables in the model. The null hypothesis can then be either accepted or rejected based on the following decision rule:

$$\text{If } F \leq F(1 - \alpha;\ p;\ n - p - 1), \text{conclude } H_0.$$
$$\text{Otherwise,} \quad \text{conclude } H_1.$$

Also *p*-values can be used in overall *F*-test analysis. Many other types of tests for the statistical significance of both the model and the coefficients are available (Kutner *et al.*, 2005). The test hypothesis may differ by context.

3.2. Model fit indicators

As mentioned earlier, the idea of regression analysis is either to describe the data or to predict the response variable values of new data. To determine the goodness of a model and to compare alternative models, an indicator is needed to tell how well a model fits the data set. For this purpose Kutner *et al*. (2005) introduce a concept of the *degree of linear association* between the response and predictor values in the model.

One commonly used indicator value for the degree of linear association is the coefficient of determination, R^2, which has values between zero and one. The value has been described as a percentage of explained variation of the model (Freedman, 2009):

$$R^2 = 1 - \frac{SSE}{SSTO},$$

where the *SSTO* represents the total uncertainty of prediction. The *SSTO* is calculated as follows:

$$SSTO = SSR + SSE = \sum (Y_i - \overline{Y})^2.$$

A high R^2 value indicates a good fit, even though a perfect model (R^2 is 1) is extremely rare (Hand *et al*., 2001). The percentage is low in cases where the fit of the model is poor, for instance, when all the predictor variables are not measured or measurable, and, therefore, not included in the model.

There are some limitations to be aware of when using R^2 (Kutner *et al*., 2005). Firstly, a high coefficient of determination does not necessarily indicate that good predictions can be made or that the estimated regression line or plane has a good fit. The prediction intervals can be very wide for the context, because only relative reduction is measured. Even though R^2 indicates a good fit, the shape of the curve can be even nonlinear. Secondly, a low coefficient of determination does not always indicate that the response variable is not related to the predictor variables; it only means that the relationship between the variables is not linear. Thirdly, an important issue to note in multiple linear regression is that by including more variables in the model, the R^2 value cannot be reduced. This may cause invalid results. One solution is to adjust R^2 by the number of predictor variables of the model:

$$Adjusted\ R^2 = 1 - \frac{SSE/(n-p+1)}{SSTO/(n-1)},$$

where both sums of squares are divided by the degrees of freedom associated with them, n being the sample size and p the number of predictors in the model. With this adjustment, adding a predictor variable may reduce the adjusted R^2 value. A big difference between R^2 and adjusted R^2 values indicates invalid results.

According to Kutner *et al.* (2005) no single descriptive measure can offer enough information on whether the model is useful for specific applications, and, therefore, descriptive measures should not be used separately. It depends on the field of study how high a measure is considered significant. According to Freedman (2009), in fields like sociology even 20 percent of explained variation might be remarkable, because of, for instance, large random effects and difficulties in measuring.

3.3. Variance inflation factor

Multicollinearity stands for predictors correlating with each other in the same model. In case of perfect correlation between predictor variables X_1 and X_2, it can be shown that an infinite number of perfectly fitting, but totally different response functions with totally different response values can be found (Kutner *et al.*, 2005). Usually predictors are not perfectly correlated, but the effects of multicollinearity still have some relevance. The main effect of multicollinearity is the impreciseness of predictions, although when new values are within the region of observations, the effect is not as drastic (Kutner *et al.*, 2005).

Multicollinearity can be detected by different methods. Big changes in regression coefficients, when a predictor is added, or deleted or a nonsignificant result from a test for individual predictor variables often indicate multicollinearity. Insensible coefficients and very wide confidence intervals often imply multicollinearity. A correlation matrix between predictor variables also shows high values for multicollinear predictors. One widely accepted method (Kutner *et al.*, 2005) for multicollinearity detection is calculating the *variance inflation factor* (*VIF*).

The variance inflation factor measures the inflation of variation in estimated regression coefficients as compared to when the predictor variables have no linear relationship between each other. The first step in calculating the variance inflation factor is to perform the least squares regression for each X_i of the model with all the other predictor variables left as predictors. The equation for $i = 1$ is

$$X_1 = \beta_0 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon,$$

where β_0 is the constant and ε is the error. The next step is to calculate the variance inflation factor for X_i:

$$VIF_i = \frac{1}{1 - R_i^2}.$$

The coefficient of determination, R_i^2, is calculated using the equation presented earlier. As a general guideline, if the largest variance inflation factor has a value over 10, the model should not be used for prediction (Kutner *et al.*, 2005).

3.4. Residuals

Residuals measure the difference between the observed and the estimated values of the response variables. Residuals are used to study different types of departures from the model. These include nonlinear relations, error terms not meeting their assumptions for linear regression modeling, and missing important predictor variables (Kutner *et al.*, 2005). Also highly influential values, *outliers*, can be detected with residuals. This will be discussed in Section 3.5.

Residuals for every data point are calculated by the following equation:

$$e_i = Y_i - \hat{Y}_i.$$

In this equation, Y_i is the observed value and $\hat{Y}_i$ the fitted value for the response variable Y. The residual is closely related to the error term of the model:

$$\varepsilon_i = Y_i - E(Y_i),$$

where $E(Y_i)$ is the expected value for Y. The difference between the residual and the error term is that e_i is considered as the observed error, while ε_i stands for the unknown true error of the model. In the regression model, the error terms are normally distributed random variables with zero mean and a constant variance.

Therefore, the distribution of the residuals should have similar properties, if the fit of the model is correct. The median is expected to be close to zero. The absolute values of both the minimum and the maximum, and the first and third quarters of residual distribution should be close to each other. Any aberration of these values should be examined closely. However, the statistics might look

reasonable even when the fitted model is completely wrong. Residual plots may offer a better insight.

Residual plots are used to illustrate the departures from the model, which they usually do more effectively than the regular scatter plots (Kutner *et al.*, 2005). This is demonstrated in Figure 3.

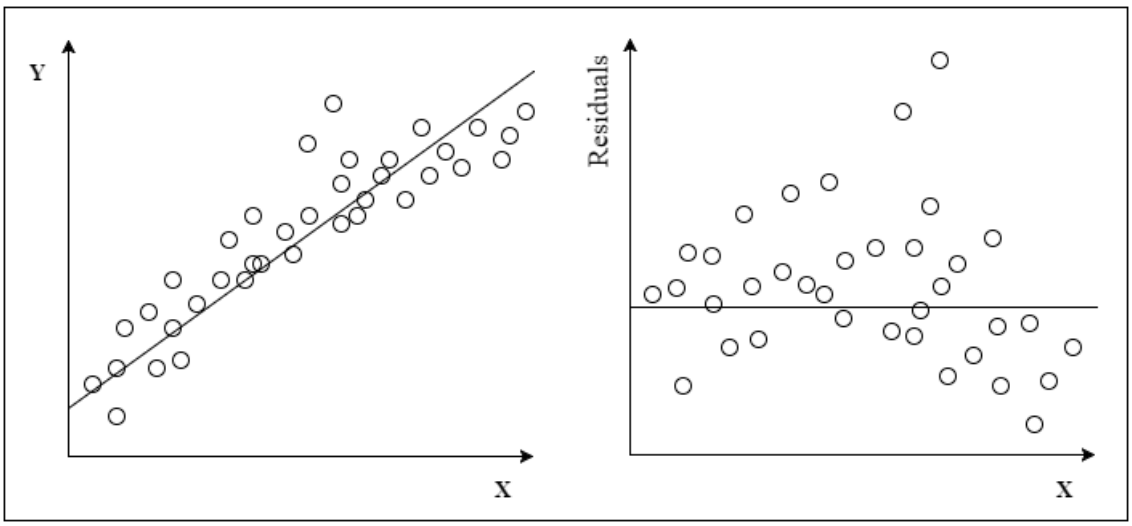

Figure 3. A scatter plot and residuals.

The left side of Figure 3 is a regular scatter plot between the response variable Y and the predictor variable *X*. The relationship between them seems quite linear. The right side of Figure 3 shows the residual plot for the same situation. In linear regression, residuals are supposed to disperse evenly along the regression line. Clearly, this is not the situation for this case since the data points form a curved pattern. A better fit and especially, better predictions would be found by fitting a curvilinear regression line.

3.5. Outlier detection

Outliers are data points located far away from other values in a data set (Maddala, 1992). The term outlier stands for cases with one or more extreme values and cases with substantially larger residuals than other cases in the model. In regression analysis the latter case is usually more influential (Bollen & Jackman, 1985). An outlier is often, but not necessary, also an *influential data point* that has a larger impact on the regression line than most other data points in the data set. Deleting an influential data point causes big changes to the model. According to Bollen and Jackman (1985) outliers may appear because of erroneous measurements, because important variables are omitted from the model or have not been measured at all, or even because they represent the natural variation in the data set. The outliers caused by the last reason usually blend into the data set, when the sample size is increased.

A regular scatter plot between the response variable and the predictor or in multiple regression the linear combination of the predictor variables can help to detect outliers, as presented in Figure 4.

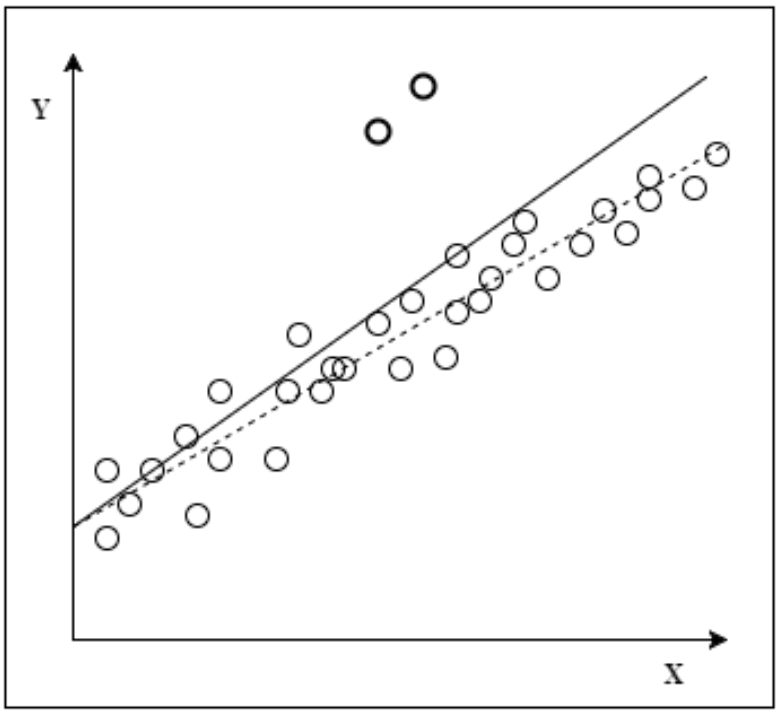

Figure 4. A scatter plot with two distinct outliers. The dashed line represents the location of the regression line after removing the outliers.

Residual plots may help to reveal less obvious outliers. Also, some distance measures have been developed for outlier detection. Bollen and Jackman (1985) emphasize that distance measures in outlier detection should not be used as a substitute, but as an aid for careful statistical analysis.

Cook's distance (Cook, 1977) measures the influence of each case to n fitted values:

$$D_i = \frac{\sum_{j=i}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{(p+1)MSE},$$

where $\hat{Y}_j$ is the fitted value for each n and $\hat{Y}_{j(i)}$ is the corresponding fitted value in which the ith case is removed from the fitting of the regression model. The factor p is the number of variables in the model and MSE is the mean squared error. For the cutoff value for recognizing highly influential data points has two different points of view: $D_i > 1$ is suggested by Cook and Weisberg (1983) and $D_i > 4 / n$ is suggested by Bollen and Jackman (1985). Stevens (1984) points out

that Cook's distance cannot always detect values as outliers, when both the response and the predictor variable have outlying values.

Mahalanobis distance (Mahalanobis, 1936) measures the distance between a multi-dimensional data point $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^{\mathrm{T}}$ and the data set distribution:

$$D_M(x_i) = \sqrt{(\boldsymbol{x}_i - \boldsymbol{\mu})^T S^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})},$$

where $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)^{\mathrm{T}}$ is the mean vector and S represents the variance-covariance matrix of the data. A large value for Mahalanobis distance indicates that the data point might be an outlier. The threshold value for outliers can be, for instance, a χ^2-test result. For reliable results, Mahalanobis distance has to be calculated by robust regression methods (Filzmoser, 2004).

4. Subset selection in regression

Given a set of X variables, the goal of subset selection in regression is to find such subset of variables that describes the variation of Y well. An ideal result is to be able to form a tight band of data points around the regression line or plane. The obvious goal is to the subset of variables that fits the data the best.

For comparing linear regression models, there are many different criteria (Kutner *et al.*, 2005) of which a few commonly used ones will be presented. Both R^2 and adjusted R^2, which were described in Section 3.2 for model fit indication, are used also as model selection criteria. Another criterion for the model selection is *Mallows' C_p*, first introduced by Mallows (1973). Mallows' C_p calculates the sum of squared error, *SSE*, for all fitted values:

$$C_p = \frac{SSE}{MSE} - n + 2(p+1).$$

The n value is the number of data points used in the model creation, and p is the number of variables of the model, the predictor variables and the intercept. Usually, a smaller Mallows' C_p means a better fit. However, $C_p < p + 1$ often indicates a positive bias, which may indicate too optimistic results. According to Daniel and Wood (1999), the best model is the one in which C_p is closest to p + 1 from above. The three criteria presented above usually give approximately the same results (Miller, 2002), but to be sure, all values should be taken into account in model selection.

One significant issue – the problem of overfitting – occurs in every model building process (Miller, 2002). Overfitting appears when the same data are used both in selecting predictor variables for the model and in estimating the regression coefficients. Basically, an overfitted model describes the current data set very well, but fails to generalize to the whole data space. This problem can be overcome by distributing data to multiple subsets to have separate samples for training and validation.

Dozens of algorithms, and variations of them, have been created for the tasks of subset generation and selection. The main factor in choosing the subset selection method is the environment, where the search will be applied in. For instance, a large amount of predictor variable candidates or complex relationships between them may have an effect on which selection methods are usable. In this chapter, a few commonly known methods for finding well-fitting subsets to be used in linear regression are looked into. Section 4.1 describes the stepwise selection methods, in which a single variable is either added to or

removed from the model in each step. Section 4.2 concentrates on exhaustive selection methods, in which all possible variable combinations are reviewed.

4.1. Stepwise selection methods

Stepwise subset selection stands for methods where the model is created step-by-step, either by adding or removing variables from the model. In general, stepwise selection methods are efficient, but not as reliable as exhaustive methods. The stepwise subset selection methods described here are forward selection, Efroymson's method, and backward elimination.

Forward selection (Miller, 2002) starts with no predictors in the model. The predictor variables are then added to the model one by one until adding a variable does not improve the model anymore or a predetermined stopping rule is met. A clear advantage of forward selection is its computational inexpensiveness. In the first step, when the model is empty, there are k calculations to be made, k being the total number of predictor variable candidates. The number of calculations decreases by one in every step. Therefore, if every variable is included in the model, the total number of evaluated subsets is $k(k+1)/2$. On the other hand, the inexpensive algorithm has its downside: finding the best fitting subset is not guaranteed (Miller, 2002), especially in cases where linear combinations of predictors have more predictive value than individual variables.

Backward elimination (Miller, 2002) is a reverse method to forward selection. At the start of the first step, all variables are included in the model. The effect of deleting each of the variables is evaluated one at a time. In least squares regression, the variable chosen for deletion is the one which leads to the lowest possible residual sum of squares for the rest of the model. The procedure is continued until the model cannot be improved anymore.

The maximum number of calculations to be made in backward elimination is the same as in forward selection. In the first step, all the explanatory variable candidates are included in the model, and the effect of excluding every variable is evaluated. Therefore, the number of evaluated subsets is k, which is the number of candidate variables. The number of calculations to be made reduces by one in every step, resulting in the same maximum of evaluated subsets as in forward selection, $k(k+1)/2$. In practice, backward elimination is often computationally far more demanding than forward selection because the models are usually kept relatively simple by including only some predictor variables (Miller, 2002): In a case of 50 candidate predictor variables of which a maximum of 10 are to be selected in the model, the first step in backward elimination includes 50 calculations, then 49, then 48, until the size of interest is

reached. In forward selection, the maximum number of calculations proceeds until 10 variables are included in the model.

In comparison of forward selection, the backward elimination method tends to leave the linear combinations correlating with the response variable to the model more often (Mantel, 1970). Still, it is possible that backward elimination cannot find the best possible subset of each size, as, for instance, Berk (1978) has demonstrated.

Efroymson's algorithm (Efroymson, 1960) is a modified version of forward selection (Miller, 2002). The first step is the same as in forward selection: The first predictor variable is added to the empty model. After an addition, it is checked, if any of the variables added earlier can be deleted without affecting the error sum of squares in the model. The variables are then deleted from and left in the model based on the calculated *R-ratios*. The procedure is continued till the model cannot be improved anymore.

The criterion for variable addition in Efroymson's algorithm is calculated with the following *R-ratio*, where p is the number of variables in the model and k is the total number of candidate variables:

$$R = \frac{SSE_p - SSE_{p+1}}{SSE_{p+1}/(k-p-2)}.$$

In the equation, two different error sums of squares are calculated: the SSE_p value is for the current subset of selected predictor variables and the SSE_{p+1} value is the smallest error sum of squares that can be found by adding a new predictor variable to the current subset. The resulting *R-ratio* is then compared to the predetermined *F-to-enter* value with the selected risk level from *F*-distribution, and if greater, the variable will be added in the model. The ratio for variable deletion in Efroymson's algorithm is calculated as follows:

$$R = \frac{SSE_{p-1} - SSE_p}{SSE_p/(k-p-1)}.$$

Again, two different error sums of squares are calculated. The SSE_p value is for the current selected subset of predictor variables, and the SSE_{p-1} value is the smallest value found by deleting one variable from the current subset. The *R*-ratio is then compared to the predetermined *F-to-delete* value. If *R* is less than *F-to-delete*, the variable is removed from the model.

The maximum number of calculations in Efroymson's algoritm is significantly higher than in forward or backward selection. The first step

includes k calculations, and the rest of the steps include $(k-p)+p=k$ calculations, ending up in a total maximum of k^2. In practice, backward elimination might still be computationally more demanding. The algorithm performs better than forward selection in cases where some of the predictor candidates are highly correlated with each other, but it is still not guaranteed to find the best fitting subset (Miller, 2002).

4.2. Exhaustive selection methods

One very straightforward technique for subset selection in regression is generating all possible subsets. Naturally, with this method finding the best fitting subset is guaranteed. The clear disadvantage of this method is cost: the number of possible subsets of k candidate variables is 2^k- 1 which means that an additional variable approximately doubles the computational cost (Miller, 2002).

There are a few ways to reduce the cost of exhaustive search without losing information. Firstly, the maximum size of a resulting subset can be limited based on framework specific knowledge: It is unlikely that one response variable has dozens of predictor variables that all significantly improve the model. Secondly, a more effective branch-and-bound search algorithm can be implemented. Miller (2002) states that a combination of these actions is usually enough for enabling the exhaustive search even in large data sets.

In a branch-and-bound search, all possible subsets of each size are divided into two branches: the ones that include the variable X_1 and the ones that do not. These branches are then divided into two subbranches including and excluding X_2 variable. This is continues, until at some point, a subset containing either X_1 or X_2 both gives a residual sum of squares a. Then suppose that the subbranch which excludes both X_1 and X_2 has a lower bound on the smallest residual sum of squares, the residual sum of squares for all the variables left, b. Now, if a is smaller than b, it can be reasoned that no subset of the same size from the subbranch without X_1 or X_2 can make a better fit than in the other subbranch. Therefore, the subbranch without X_1 or X_2 can be excluded from the search. Even more exclusions can be made, if the task is to find the best-fitting subsets of all sizes (Miller, 2002). Dividing the branches to two subbranches continues till every subset is either calculated or excluded from the search.

The branch-and-bound method has the advantages of generating all subsets, because the best-fitting subset of each size is guaranteed to be found. The amount of computation can still be substantially reduced, especially in environments, where one or more variables are clearly dominant. Kariwala *et*

al. (2013) have recently introduced a pruning algorithm for speeding up the branch-and-bound method even more.

5. Algorithm requirements and preparatory work

The concept of dimensioning stands for defining the minimum capacity requirements of a network or a network management system (Section 2.3). The goal of this study is to develop a new dimensioning modeling algorithm both to overcome the problems arisen with the earlier dimensioning models and to respond to future challenges. The main change from the previous dimensioning solutions is that instead of predetermined models, the new dimensioning modeling algorithm will generate models based on actual performance data. As earlier, the usage of every system resource or other application performance attribute will be modeled separately. The difference in the current approach is that the best load profile for every resource and other performance attributes will be searched from the space of all possible subsets of predictor variables. The resulting algorithm is described in Chapter 7.

The network management systems operate in very complex environments which influences the dimensioning of the system. First of all, every customer has a different setup and a combination of external systems sharing the same hardware resources. Thus, it is important to collect enough data from different environments to ensure that the data describe the whole range of the resource usage. Secondly, new components are added with new releases. Naturally, the resource usage data of the new components is not available from customer environments before the release. Thirdly, it is very common that all of the predictor variables are not measured. The unmeasured variables can sometimes be seen as a source of outlying values in an otherwise well-behaving data set. These characteristics of the data do not necessarily cause problems, if they are acknowledged and handled properly.

The first network management tool from Nokia was released in early 1990s. Tremendous changes have taken place after the first release, in both networks and network management. Therefore, also the network environment has become more and more complex during past years. At the same time, customers have started demanding more accurate predictions of resource usage and other performance indicators of interest as well as more tailored networks. Thus, understanding the performance of both networks and network management systems has become harder and more essential. The current dimensioning tool works with the complicated environment, but requires a lot of maintenance and testing to adapt to constant changes.

In future, the differences in customer setups are likely to rise substantially, when the products move to cloud environment and the predefined configurations can be replaced with more exact, customer-specific

configurations. The need for more specific dimensioning is contributed by the customers of the network management systems, who are currently struggling with the economic demands of more and more complex systems, and start to appreciate both the efficient usage of current hardware and the possibility of re-using old hardware. To answer this configuration need, the rough classification must be transformed to an elaborate regression. The new way of dimensioning is required both to overcome the problems of earlier dimensioning solutions and to prepare for future ideals.

In this chapter, the requirements for the dimensioning of Nokia's network management system (Section 5.1) are looked into. The available performance data are presented in Section 5.2 and the necessary pre-processing are described in Section 5.3. Finally, Section 5.4 looks into the challenge of selecting a suitable data analysis tool and presents R, the selected environment for statistical computing.

5.1. Algorithm requirements

The purpose of the dimensioning modeling algorithm is to find the best combination of variables to predict the usage of a hardware resource or other application performance attribute by using the multiple linear regression. To achieve rational and trustworthy predictions, some requirements for this model creation process must be stated. First of all, the variable selection must be data-driven. Secondly, the models must be kept relatively simple. Thirdly, the models created by the algorithm must be reliable.

Data-driven subset selection has a few advantages over using predetermined combinations of the variables. In the network management system framework, the resource usage varies between different customers, since every user can have a unique combination of internal and external applications. Resource usage is affected also when the system is updated and when more resources are added to the system. This makes the maintaining and updating of the models challenging and requires a lot of knowledge of the particular system when the predictor variables and factors are determined manually. The data-driven model generation method eases up the process substantially. Also, the possibility of human error is decreased and unknown relationships can be discovered. However, to achieve reliable results, a representative sample with enough data points must be available. It is important to note that the algorithm does not distinguish whether the relationship it finds between the variables is causal (Hand *et al.*, 2001).

The simplicity of the model is important mainly to keep the model understandable. Being able to understand how the model is constructed is

important to the engineers analyzing the network management system performance. A substantial challenge is to find a good balance between the accuracy and complexity of the model, since increasing one tends to decrease the other (Hand *et al.*, 2001). Therefore, the question to ask is how much more of the variation has to be explained by the predictor variable candidate for it to be worth to be included in the model. In addition, the computational cost of the modeling must be considered. The implemented model generating algorithm emphasizes understandability: It selects simpler models over more complicated ones, and the maximum number of predictor variables can be limited, when the best subsets are generated.

Generally, for a multiple linear regression model to be reliable it has to meet specified standards (Kutner *et al.*, 2005). An algorithm meeting the requirements will generate the best possible subsets for predictive analysis. All the necessary statistics, which were described in Chapter 4, must be calculated in the model generation process, and thresholds for rejecting the model must be defined. The possibility for overfitting must be managed with external validation. Overlooking the reliability requirement can, especially in predictive analysis, cause serious consequences, such as very misleading predictions.

5.2. Data features

When generating models for resource usage estimation, some important issues must be considered. Firstly, representative data from the network management systems is needed. Secondly, to achieve generalizable results, it is important to have a large amount of data with enough variation. For these purposes data were collected from test environments, where simulators generate synthetic data which imitate the real network management system environment data. The test environment data are suitable, however, only for creating general guidelines, since every network has a different setup with a different combination of internal and external elements. Therefore, definite results cannot be achieved, until sufficient data from multiple different customer environments is available.

In this environment, all the dimensioning variables can be divided into two classes: load variables and processing capability variables. One load variable represents one type of load the system has to manage with, for instance, load from application or user. Processing capability variables present the usage of both the hardware and software resources. Load variables are considered as predictors, while processing capability variables represent the response variables to be estimated. The variables are numerical and continuous, that is, the variables are measured either at interval or ratio level (Hand *et al.*, 2001).

Some of the requirements for data suitable to be analyzed with multiple linear regression were discussed in Chapter 3. These were linearity between the response variable and predictor variables, no highly correlating predictor variables in a model and no highly influential data points. The linearity requirement is easily achieved in this framework, because up to the theoretical maximum capacity of a resource, a processing capability variable, the growth in resource usage is linear, if all the affecting factors are included in the model. In dimensioning, the area of interest is before the theoretical maximum value which is never to be exceed, so the requirement for linearity is met.

The second requirement is not as easily achieved. In this environment, there are often multiple variables explaining almost or exactly the same variation of the data. For instance, in the load profile of performance management data, a load amount can be measured both in files and in counters. These two do not necessarily have to be multiplications of each other, but usually at least the correlation between them is very high. Luckily highly correlated predictors are easy to detect either in the pre-processing or in the model creation phase.

The final requirement for the data can be achieved relatively easily. An environment specific feature is that if all the affecting factors are included in the model, no highly influential values appear, because the data are synthetic. However, often occasional peaks in resource usage are due to processes which are not measured or even measurable. Since the cause of a peak in this environment is typically known, the influential data point can be removed from the sample without consequences. Therefore, all the requirements for multiple linear regression can be met, it can be said that multiple linear regression is a suitable model for this environment.

The most representative feature of the test environment data is that the values of the predictor variables are decided by testers, since each test has its predetermined characteristics in the load profile. These predictor variables naturally affect the response variable, the resource usage, which is to be estimated. The variation in resource usage within a test is usually low. Therefore, a single test round, a time frame for one simulator test to run with the predetermined load profile, appears as a cluster in the data, which is illustrated in Figure 5.

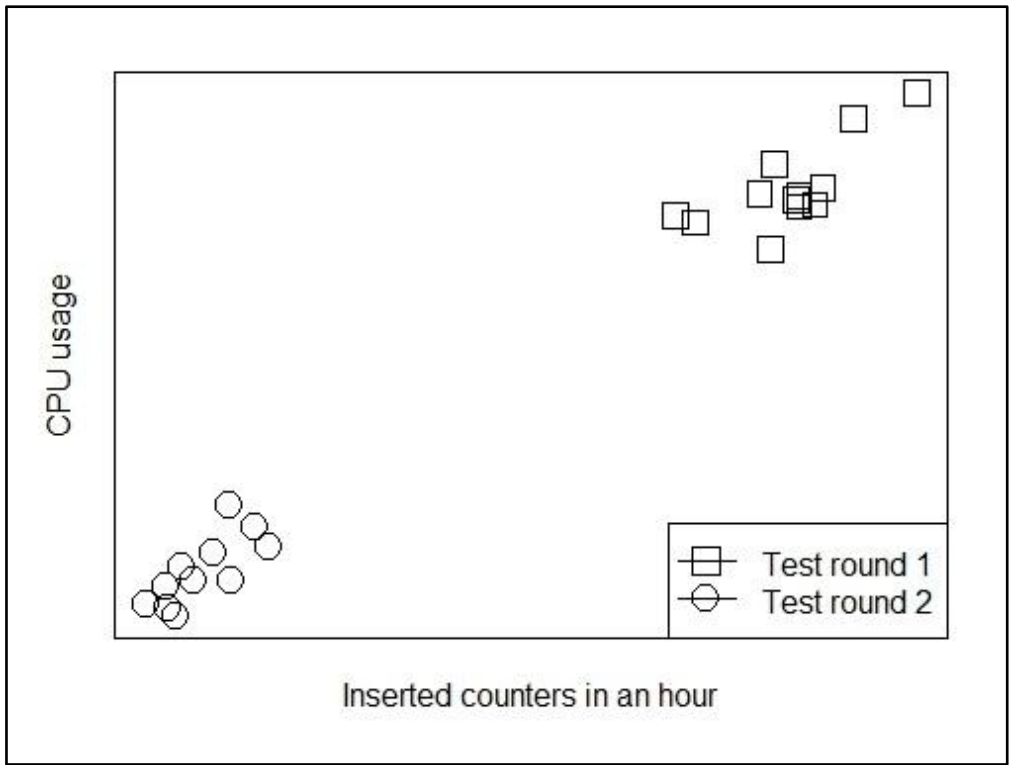

Figure 5. The test round 1 was operated with one third of maximum inserted counters in an hour and the test round 2 had the maximum load. Other CPU usage affecting factors were kept steady.

When more and more test rounds with different load profiles are run, the clusters are slowly absorbed in the data mass. Having this many test rounds is rarely possible for all the measured resources. However, linear regression analysis can be applied with data even from fewer test rounds, because the growth in resource usage tends to be linear to the theoretical maximum capacity of the resource. In this approach, the empty space between the clusters is assumed to have linearly growing values with the same variance as the executed test rounds.

Data from customer's environment do not have a lot of variation, if the number of integrated network elements and allocated hardware resources stays the same. Therefore, data from a single customer act similarly as the data from separate test rounds, forming clusters in the data set. As with test round data, the space without data points in between the clusters would not be empty, if a complete set of data were available.

The data were exported mainly from a system monitoring device. This data was highly granular: Almost all metrics are in separate files. One file includes the same measurement from all the resources, for instance, virtual machines, in which the measurement has values. In general, the unique key of a data point in an export file is the combination of the timestamp, the metric, the resource and the record. Some files may have missing data in some of the fields. The

timestamp represents the time when the metric, either a response or a predictor variable to be used in dimensioning, was imported to the system monitoring device, not the time of the actual event. The ongoing goal is to collect all metrics in one database to ease up the mapping process for dimensioning and other purposes. Table 1 demonstrates the resulting form of data, which still requires processing before the analysis can be performed.

metric	*resource*	*timestamp*	*value*
metric1	resource1	6/2/2015 17:00	980
metric2	resource1	6/2/2015 17:13	52
metric2	resource2	6/2/2015 17:13	6724
metric1	resource1	6/2/2015 17:28	1020
metric3	resource2	6/2/2015 17:42	102324000
metric2	resource2	6/2/2015 17:50	9820
metric1	resource1	6/2/2015 18:02	1000
metric2	resource1	6/2/2015 18:15	73
metric2	resource2	6/2/2015 18:15	8901
metric1	resource1	6/2/2015 18:33	1230
metric3	resource2	6/2/2015 18:38	164903857
metric2	resource2	6/2/2015 18:50	10026

Table 1. A simplified example the initial form of data.

5.3. Pre-processing

Due to the granularity of data collected from the system monitoring device, a slightly simplified version of the actual pre-processing procedure will be presented. For the dimensioning modeling algorithm, the variables are aggregated into one hour level, since some metrics only get values once in an hour, to be able to map the actions together and to find correlations. Table 2 presents the suitable data format for the algorithm; only the separation of the response and predictor variables has to be done afterwards.

timestamp	*metric1. resource1. mean*	*metric2. resource1*	*metric2. resource2. sum*	*metric3. resource2*
6/2/2015 17:00	1000	52	16544	102324000
6/2/2015 18:00	1115	73	18927	164903857

Table 2. Two aggregated data points.

As the response variable names in Table 2 show, the hourly arithmetic mean of *metric1* is calculated for the *resource1*. For some variables the total number of measurements in an hour is summed up, as it can be seen in case of the *metric2*

for *resource2* in Table 2. The metric type determines whether the mean, sum, or other statistic must be calculated for the table of aggregated values. The timestamp attribute in Table 2 is not provided to the dimensioning modeling algorithm, but presents how the variables are mapped together in the model creation phase. The algorithm does not perform mapping, but relies on the order of the parameters.

One way to transform the data from the form presented in Table 1 to a suitable form for dimensioning algorithm modeling as presented in Table 2 is to first create columns for every unique combination of the metric and the resource. In practice, even more definitive columns may appear in the raw data table. The values under the columns are from the *value* field (Table 1). Since in this pre-processing step the table still has a row for every timestamp, there are a lot of empty cells, as can be seen in Table 3.

timestamp	*metric1. resource1*	*metric2. resource1*	*metric2. resource2*	*metric3. resource2*
6/2/2015 17:00	980			
6/2/2015 17:13		52	6724	
6/2/2015 17:28	1020			
6/2/2015 17:42				102324000
6/2/2015 17:50			9820	
6/2/2015 18:02	1000			
6/2/2015 18:15		73	8901	
6/2/2015 18:33	1230			
6/2/2015 18:38				164903857
6/2/2015 18:50			10026	

Table 3. The data form after the first pre-processing step.

The next pre-processing step is to specify an aggregate value for the hourly representation of each variable. Aggregating is risky, because processing an event sometimes continues to the next time slot. However, in one-hour level the risk is minimal, since events are usually scheduled to the start of the hour. Transforming data to one-hour level also requires some environment-specific knowledge for understanding which statistics can be used without losing important information. In practice, this step can be easier to implement after separating the columns to groups by indicator values. Environment-specific knowledge is also needed when the predictor variables must be found amongst the response variables.

The granularity of the data can be dealt with in a few ways. One option is to perform the steps separately in all files and afterwards merge the tables based

on the timestamp. This requires more work in the first step, but the second step is easier to manage. The tables can also be merged before the first step, then performing the first step is fast but the second step is more complex.

The idea behind the pre-processing procedure is quite simple, but in practice it includes a lot of time consuming work. Therefore, once the pre-processing steps where determined and the requirements for the data became clear, the procedure was automated. Before that all the work was done manually and even more work was required due to the granularity and incompleteness of the data and mapping issues. Even though the unified data form and automation substantially reduced amount of manual pre-processing, some work is still required in addition to the described process. In total, the pre-processing of the data has taken dozens of hours.

When a column is created for every unique combination of certain definitive column in original table, a combinatory explosion is a risk to consider. Luckily, the majority of the columns are for response variables which cannot explain each other. The amount of predictor variable candidates can be limited to those which affect either the same resource or the whole system. With these limitations, the size of the data matrix for a single response variable is likely to be closer to dozens of columns instead of hundreds.

However, limiting the number of variables used in subset selection algorithms may still sometimes be necessary, especially if an exhaustive search has to be performed. There are some approaches that can be applied in this situation. First of all, when there are highly correlating predictor variable candidates, the ones with less correlation with the response variable can be removed, as mentioned above. Secondly, linear relationships between certain response and predictor variables are very unlikely. Still, in complex systems surprising elements can have an effect on each other.

Naturally, when excluding the variables from calculation, the best possible subset is not guaranteed to be found. Removing variables is always risky, but if the environment is well understood by engineers, the number of predictor variable candidates can be limited. It should be noted that if computational cost is not an issue, there is no reason to rule any variables out before the subset selection procedure, since the variance inflation factor detects multicollinearity more reliably.

5.4. Tools for data analysis

Most of the requirements for the data analysis tool were not fully known at the start of the project, because the prediction model was yet to be decided. Only the need of some basic functionality, like reading data from and writing to

comma-separated value format and databases, were specified beforehand. It later became obvious that a popular and highly extensible tool having the basic statistical methods as well as data mining and data analysis techniques was needed. Therefore, R, an open source language and environment for statistical computing and graphics (Ihaka & Gentleman, 1996; R Core Team, 2015) was chosen. R is available under the terms of GNU General Public License (Free Software Foundation Inc., 1991). R offers a large variety of functions for statistics, data analytics, and data mining. The comprehensive R archive network repository alone includes nearly 7000 packages for extending the R base functionality (The R Foundation, 2015). For computationally more demanding tasks, functions from traditional programming languages, such as C and C++, can be utilized.

R has been described as "a dynamic, lazy, functional and object-oriented programming language with a rather unusual combination of features" (Morandat *et al.*, 2012). The R language is heavily influenced by S and Scheme languages (Ihaka & Gentleman, 1996). The main differences between R and S languages are, according to Morandat *et al*. (2012), the open source nature, better performance, lexical scoping and garbage collection of R. Being a combination of multiple different languages is an advantage, when useful features from different languages are combined (Ihaka & Gentleman, 1996), and a disadvantage when considering its computational abilities (Morandat *et al.*, 2012).

The R language is not the most efficient programming language, but its performance can be substantially increased by recognizing the bottlenecks and acting on them. Visser *et al.* (2015) have studied the subject from computational biology point of view, but the same ideas also apply to the network dimensioning field of study. Inefficient R-style loops can be replaced with vectorized C-style loops, dynamic data structures with oversized static versions, and the most time-consuming functions can be implemented in C code. R-functions frequently contain additional features which are often useful, but slow down the computation. If these features are not needed, implementing simple functions from scratch can reduce the calculation time. According to Visser *et al.* (2015), tremendous improvements in performance can be made by implementing these changes.

It can be concluded that the basics of the R language are easy to learn, especially by computer scientists, but it has a steeper learning curve for efficient programming. This is one of the reasons why R was chosen to be used for this project: Even though implementing the functions requires more knowledge of the system, reviewing and understanding the resulting program for later

purposes is quite easy. Other factors affecting the decision were the high accessibility due to the open source nature of R, the extreme flexibility of a Turing-complete programming language and easy solutions to almost any problem being available through the R-base or external packages. In conclusion, the requirements for the data analysis tool for this task were met.

Another tool was considered to be used side by side with R. The purpose of this was to ease up the visualization and graphing process for employees without R-knowledge or time to learn it. Naturally, when a graphical user interface is added, flexibility is compromised. The goal was therefore to find a good combination of flexible function implementation possibilities and easy-to-use visualization tools. Some options considered were Tableau (Tableau software, 2003) and Alteryx (Alteryx Inc., 2010), which both have an intuitive user interface and a possibility to implement user's own R functions. The final choice has not been made yet, and is not in the scope of this study.

6. Results

The framework consists of dozens of response variables and predictor variable candidates. The purpose of the model generator is to be able to create and maintain trustworthy models for predicting the values for all the response variables with least human effort. However, a completely automated solution is likely to be too risky, because there is no way to assure that the sample is a good representation of the data. This is related to the general uncertainty in predictive analysis: The reality is not perfectly regular, and, therefore, no perfect models exist. Obviously, a good balance between automation and human effort has to be found to gain both an effortless and a trustable model.

Multiple linear regression is a straightforward and traditional data analysis method. Several factors speak for its advantages in the network management system dimensioning. Firstly, the multiple linear regression with predetermined predictor variables has been successfully used before with earlier versions of Nokia's network management system. This is due to the mostly linear relationships between the response and the predictor variables. However, if at some point it is found out that not all of the response variables have linear relationships with their predictors, the modeling process can be expanded to, for instance, logistic, logarithmic or curvilinear relationships. This leads to the second advantage of multiple linear regression, flexibility. Thirdly, the linear regression model is simple. This is important for achieving low computational cost, but also for understandability. Intelligible model ensures that its quality can be monitored with little effort and the effect of a single predictor variable can be detected.

The subset selection algorithm for linear regression generates the best possible subsets with an exhaustive branch-and-bound search as described in Section 4.2. The reason for selecting an exhaustive search over stepwise methods is straightforward: The best possible subset is not guaranteed to be found with stepwise methods in this framework, because the relationships between the candidate predictor variables are manifold. From exhaustive search methods, branch-and-bound was an obvious choice, because it always allows the best result to be found, but with often requires significantly less computation (Miller, 2002). Moreover, if the number of predictor variable candidates exceeds the limit of variables that can be processed, a stepwise backward selection method can be used in preselecting variables for the branch-and-bound method.

The generated models are evaluated with various tests of significance, model fit indicators, and variance inflation factors to be able to reject the

completely untrustworthy models and to rank the models from the best to the poorest. To avoid overfitting in subset selection, a 4-fold cross-validation is applied for the model fit indicator and residual standard error calculations. The algorithm returns all the ranked models that are not rejected based on the statistics so that the user has room for adjustments with several models to choose from instead of only one. Finally, the outlier detection procedure is applied, after which the models are ready to use.

The running time of the algorithm depends on the number of predictor variable candidates and on the number of data points. A test with a very large number of both, predictor variable candidates and data points was run in few minutes, but a regular case with less than twenty predictor variable candidates was run in approximately fifteen seconds. Overall, even with the exhaustive branch-and-bound search the algorithm can be run in very modest time.

The structure of the predictive model generation algorithm is previewed in this chapter. Section 6.1 gives a deeper insight in the input and the output of the algorithm and defines guidelines for output analysis. The external validation method, 4-fold cross-validation, is presented in Section 6.2. The outlier detection process to be applied is discussed in detail in Section 6.3. Finally, Section 6.4 describes the visualization and the usage of the models generated by the automated algorithm.

6.1. Input and output

The suitable input form for the algorithm was presented in Table 3. The response variable Y is passed to the algorithm as a vector. The predictor variable candidates, X's, are passed in a data matrix. The data types chosen for the input are the simplest and the most efficient ones to process in R (Visser *et al.*, 2015). The options were defined by the ready-made functions used in the dimensioning modeling algorithm. As mentioned in Section 5.3, pre-processing is needed to achieve the required input form.

Table 4 presents the data form for the output of the algorithm. The name of the resource, the rank of the model, model fit indicators and other statistics are presented in the output data frame along with multiple linear regression models. One row in the output data represents one model: The one ranked the first is the best model for estimating the resource according to Mallows' C_p.

Resource	*Rank*	*Mallows' C_p*	R^2	*Adjusted R^2*	*F-statistic*	*p-value*
Y	1	2.89	0.80	0.6599	29.22	< 0.001
Y	2	4.02	0.63	0.63	15.28	0.0358
Y	3	4.22	0.61	0.61	24.13	0.0021
Y	4	5.01	0.59	0.57	10.02	0.0411
Y	5	5.38	0.59	0.57	11.99	0.0497
Degrees of freedom	*Residual mean square*	*Intercept*	*Weight X_1*	*Weight X_2*	…	*Weight X_n*
95	1.62	100	0.33	0.2	…	0
100	1.68	150	0	0	…	0.5
100	1.77	160	0.1	0	…	0.42
95	6.84	90	0.35	0.25	…	0
100	12.06	75	0.15	0	…	0.2

Table 4. Output data format.

The first column, *resource*, presents the name of the resource for which the model has been calculated. At this point, the *rank* variable ranks the model based on the *Mallows' C_p* variable. Still, to keep the model rank understandable and to allow the possibility for more intricate model ranking later on, both variables are included in the output separately. The R^2 and *adjusted R^2*, *F-statistic* and *p-value* are the indicator values for the fit of a model, as presented in Chapter 4. The *residual mean square* is the standard deviation for the residuals. Finally, the *degrees of freedom* indicate the number of data points used in the model calculation minus the number of parameters in the model.

The first parameter of a model is the *intercept*, also known as β_0, which was described in Chapter 3. All the predictor variable candidates have a weight $(\beta_1, \ldots, \beta_p)$ assigned for each model. If the variable is not included in the model, the weight is zero.

All the statistics describe slightly different aspects of the model allowing the combination of all the values to be taken into consideration instead of relying on a single statistic. The limits for good, average or poor models can be adjusted, when more data are available and data features are better known.

6.2. External validation

The problem of overfitting was briefly mentioned in Chapter 4: Using the same data for both subset selection and model training is likely to produce models that do not generalize to the whole data space when only a limited sample is available (Arlot & Celisse, 2010; Hand *et al.*, 2001). This basically means that the

model describes the current data very well, but is less useful when predicting new values.

In predictive analysis, a model can be evaluated with a score function either internally or externally (Hand *et al.*, 2001). In internal evaluation, the same data are used both for training the model and evaluating it. One score function commonly used for internal evaluation in linear regression is *Mallows' C_p*, which was described in Chapter 4. In external validation, the data are split in to two or more exclusive samples: One or more for training, and one for validation. A model is generated from each of the training samples, while the score function for the model is calculated from the evaluation sample. Due to the independence of the selected samples, the score function value from the validation data is unbiased.

The simplest external validation method is *holdout*, in which the data are divided into two separate sets, the training sample and the validation sample. The holdout method is computationally convenient, but the evaluation result might depend heavily on which data points end up in which data set. *K-fold cross-validation* is an improved version of the holdout method in which the data are divided into *k* data sets, and the training and evaluation is repeated *k* times, and the statistics for the score function value are calculated (Kohavi, 1995). In the k-fold cross-validation method, the impact of the data partitioning decreases but *k* times more computation is needed than in the holdout method.

Hand *et al.* (2001) state that the data and the features of the problem should always have an effect in choosing a method for the model evaluation. In the framework of this study, some factors must be taken into account. Firstly, a computationally efficient method is necessary, especially because an exhaustive branch-and-bound search through the space of possible models is performed. Secondly, a lot of different data are collected which decreases the probability of selecting unbalanced data sets. As an external evaluation solution, k-fold cross-validation was preferred over the holdout method in this study to reduce the bias in results. However, the value for k was kept relatively small, because of the computational cost.

In this study, the *k* was set to four. The external validation process for the model started with a random division of the data to four equal-sized sets. The data from three of the sets was used for training the model and the one was used for validation. This was repeated until each of the sets was used for validation exactly once. The less biased validation results were then used for the model evaluation.

6.3. Outlier detection

As mentioned in Section 3.5, the distance measures for outlier detection should not be used blindly, but as aids when trying to recognize highly influential data points. Therefore, a fully automated outlier detection method was not implemented. On the other hand, leaving the whole outlier detection process to the user is also risky. In search of a well-fitting model, even the data points representing the natural variation of data might be removed by the user, if no limits are set. Therefore, a combination of automated and manual outlier detection and removal was employed. The procedure consists of three phases: Calculating Mahalanobis distance for each data point, visualizing the outlier candidates to the user to decide if they should be removed, and reconstructing the model based on the user's decision.

Since Mahalanobis distance has performed well in test environments (Oyeyemi *et al.*, 2015), it was chosen over Cook's distance and other options. Mahalanobis distance is calculated separately for every data point, and the values are compared to the χ^2-test result. The values that do not follow the χ^2-distribution with a five-percent level of significance are presented to the user as outlier candidates.

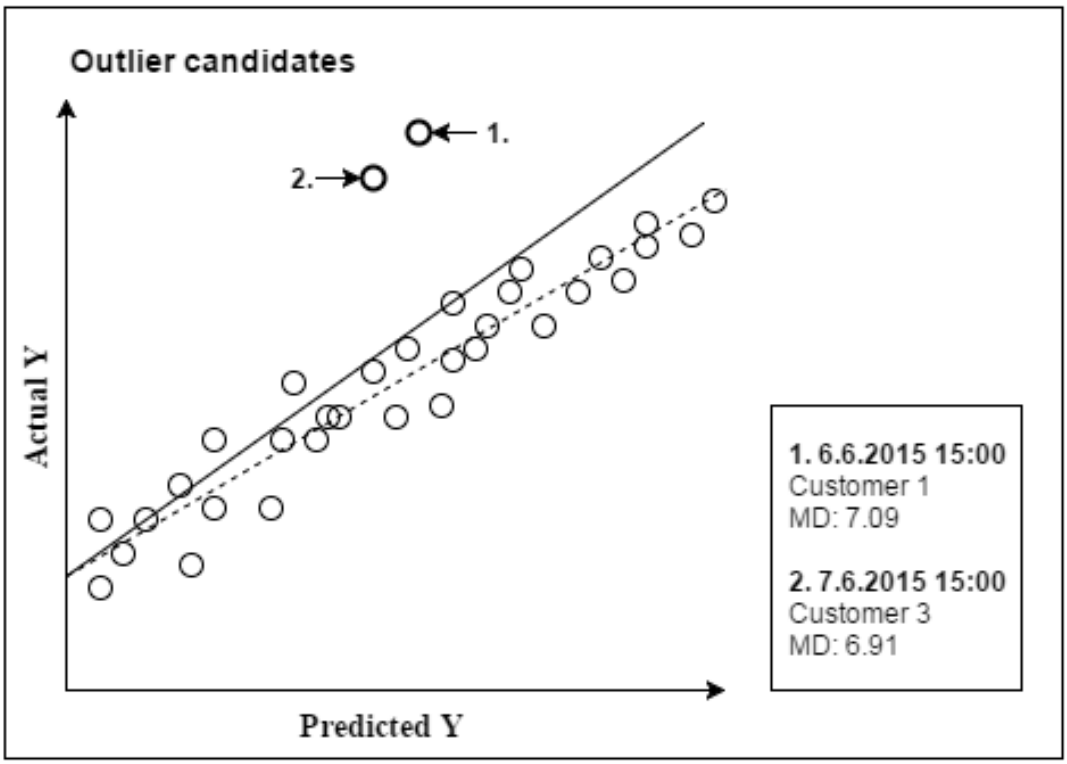

Figure 6. Outlier candidates presented to the user.

The outlier candidates are visualized with a scatter plot, where the actual values of the response variable are presented in the Y axis, and the predicted

values for the response variable are presented in the X axis (Figure 6). The predicted values are calculated from the linear combinations of the predictor variables, therefore flattening the multidimensional data to two dimensions. In an ideal situation the data points follow the regression line closely. However, distinct outliers tend to pull the regression line towards them which causes the model to describe the rest of the data less accurately, as illustrated in Figure 6, where the numbered data points represent the outlier candidates. The dashed line represents the placement of the regression line without the outliers.

The user must decide in the situation represented in Figure 6 whether the outlier candidates, the data points 1 and 2, should be removed from the data. Since the decision will affect the accuracy of future predictions, all possible information to aid in decision making is provided. A general guideline in this framework is that outliers affected by known, but unmeasured factors can safely be deleted from the data set. Therefore, the legend, as presented in Figure 6, shows the timestamp and the source of the data point, for instance a customer or a simulator. Based on this information the user can track down the events of the particular system at the particular moment. If there are exceptional events that would likely affect the measured resource, the outlier can be removed reasonably safely. Also the value of Mahalanobis distance (MD) is provided for enabling the comparison of the effects of different outliers.

Finally, if the user has decided to remove outliers from the data set based on the information provided by Mahalanobis distance measure, the model must be recalculated. The original model creation algorithm is employed again, and the resulting model is then checked for outliers. This is continued until no more outliers are found by Mahalanobis distance or the user decides that no more data points need to be removed.

6.4. Visualization

A linear regression model with only one predictor variable is easy to visualize with a scatter plot graph in which the response variable is in the Y axis and the predictor variable is in the *X* axis and the model is represented as a regression line, as demonstrated in Figure 2. Three or more dimensions are difficult or impossible to present with a scatter plot. Some possibilities are to present the *X* axis as the linear combination of all the parameters, which flattens the scatter plot to two dimensions, or to use the residual plot to present the distances between the data points and the regression line (Kutner *et al.*, 2005). However, these visualization methods were too complicated to be easily understood by people unfamiliar with multiple linear regression modeling. Therefore, only the

current and predicted distributions of the response variable are visualized with box plot graphs, as presented in Figure 7.

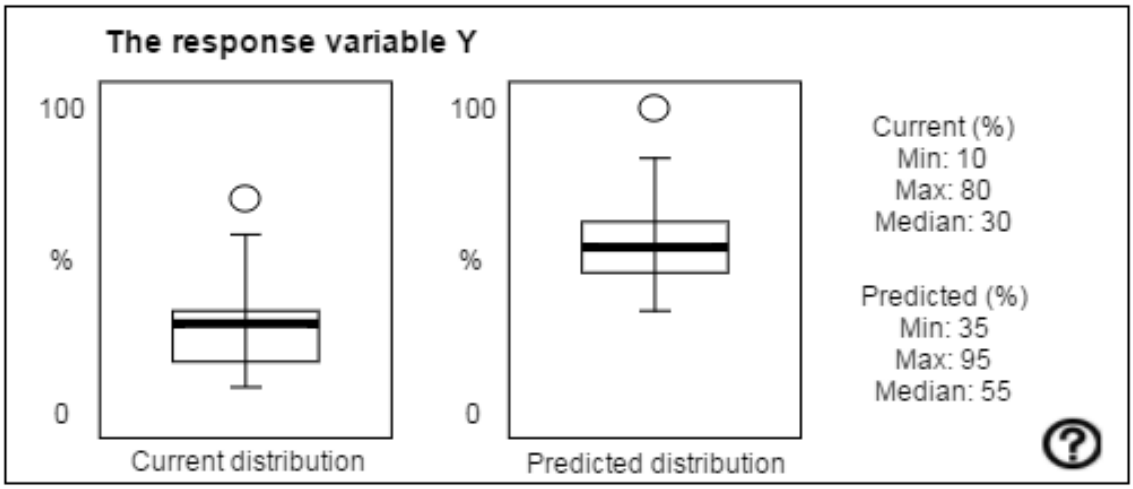

Figure 7. The realized and predicted distributions for the response variable values.

Some statistics from both distributions are offered next to the graphs for analyzing the difference in resource usage with different loads. Since a box plot graph might not be familiar to users, a description is shown by clicking the question mark symbol at the bottom right corner.

Figure 8 presents how the model is constructed and offers an option to adjust the predictor variable values for generating predicted distributions. Under the *Predictor variables* heading the predictor variables and their weights are listed. For each predictor variable a few statistics, the minimum, maximum, and median, of their current and currently predicted distributions are shown. At the start, the predicted distribution has the same values as the current distribution. The values for the predicted distribution can be changed from the text entry boxes on the right side of the screen. The circled toggle signs determine if the values are increased or decreased. The new values for the predicted distribution are updated dynamically, as also the graph presented in Figure 7.

		Current distribution			Predicted distribution			
	Predictor variables	Min	Max	Med	Min	Max	Med	
1.05 *	Predictor variable X_1	2	200	25	202	400	225	⊕ 200
0.03 *	Predictor variable X_2	122	496	301	1	375	80	⊖ 121
0.89 *	Predictor variable X_3	898	923	904	898	923	904	⊕ 0
0.67 *	Predictor variable X_4	1002	2067	1718	1602	2667	2318	⊕ 600

Figure 8. Additional information on the model and the current and predicted distributions.

These visualizations are needed to understand and to exploit the model better. Predicting how much resources are needed when the predictor variable load is either increased or decreased is the main idea in the network management system dimensioning. The maximum values in the box plots presented in Figure 7 show the current absolute maximum value for the resource. However, usually only a portion, for instance 80 percent, of the resource can be predictably used. This theoretical maximum can occasionally be exceeded without problems. Therefore, these visualizations can help determining whether a large enough portion of the distribution stays in the safe zone or if more resources are needed.

7. Discussion

Analysing the results achieved by the model generator algorithm is difficult, because the work is still at an early state: This thesis can be considered as the first step to the right direction. The main reason for this is that not enough representative data have been available from the start. In addition, new issues have appeared during the research, as usual. A lot is still to be considered and some decisions may need to be revised.

As stated in Section 5.1, the requirements for the algorithm were that it must be data-driven and the models created by it must be simple and reliable. Only automated search techniques were considered in solving the subset selection problem because of the first requirement. As one of the simplest data analysis methods, the chosen model, multiple linear regression, meets the second requirement.

The last requirement, the reliability of a model, is not as straightforward as the first two. Firstly, measuring the reliability is tricky since no statistic alone can show whether the model is reliable. A good option is to consider many different measures and weight models based on them. Secondly, analyzing the reliability of a model cannot be properly done without enough data available. Therefore, a more thorough reliability analysis cannot be done until more data are available.

At this point of the work, it appears that the algorithm has been fairly successful. However, many things can still change before the project is finished. The implementation of the algorithm still lacks some data-dependent components and the model and data maintenance are unfinished. Also, some additional functionality, which is discussed in Chapter 8, may be considered necessary or useful later on.

In Section 7.1, the model and data maintenance, as well as adaptation to a certain environment are discussed. Different solutions are presented and discussed. In Section 7.2, a comparison between the new solution for dimensioning and the preceding one is presented. Comments from the developer of the previous dimensioning tool are also included. Finally, for simulating the functioning of the algorithm, a test run with sample data is performed in Section 7.3.

7.1. Model maintenance and specification

The data analysis process does not end once the models are created. For achieving good predictions during the life cycle of a model, maintaining and updating of data and the models are necessary (Hand *et al.*, 2001). In this

environment, the data set changes all along, when new data points are constantly added, while some older data points become non-relevant. The challenge is to keep the data representative enough without losing important information. Bigger changes come with new releases, because new variables, both response and predictor, can be added. For new response variables a model must be created once enough data are available just as in the initial model creation phase. With new predictor variables all the models must be updated to ensure that all the predictor variables are included.

The data set can be kept up-to-date, for instance, simply by removing the oldest data points while adding new ones. However, the implementation of the rotation is not straightforward, because enough variation both in time and values must be ensured. Also, when updates of the network management system are released, data from the latest update must be preferred. A data set in which the rotation is based on only few factors is likely to be biased. For instance, when selecting only a certain time period, the models might be created from biased sample. Therefore, minimum values for the number of data points, variation, and time period must be specified.

Refreshing data and updating models often is computationally demanding and does not offer substantial improvements to the predictions. On the other hand, if the data set is refreshed rarely, the models will be outdated. One option is to refresh data regularly after a fixed time period, for instance, a week. Another, maybe computationally a less demanding way is to run the rotation after significant changes are made in data, such as software updates. However, automation for the latter option is quite a challenging task.

The number of predictor variables may increase, or occasionally decrease, with new releases. Unless data are updated, this leads to a situation, where values for the new predictor variables are missing from the old data, and, therefore, the old data become outdated. In such a case, data collection can be started again, or the value for the new predictor variable in the old data can be set to zero. In the latter option, more data for model creation are available, but the new predictor variables do not have enough variation for model creation. In practice, a combined solution between the two options may be preferable, but evaluation is needed for finding the best solution.

Maintaining and updating the models properly may not be enough for trustworthy predictions. Sometimes specialization for a particular customer environment is necessary. A common situation in customer environments is that third-party software has been included in the network management system. Commonly, the input load is not calculated from these elements. This

causes an unexplained change in the values of affected response variables, and the measures move further away from the regression line.

One solution to this problem is collecting data from the particular customer environment and creating personalized models. The problem with this approach is that in one environment the data do not often have much variation, because the values for predictor variables tend to move only within certain limits. Therefore, developing the customer-specific models is likely to lead to biased predictions. However, if enough variation within the environment is available, this method could work very well.

Another solution is to maintain a database table for the distances between the predicted and actual values. If the model fits the data, the mean for these values from a longer time period should stay close to zero. However, if the distances seem to be constantly large, or a repeating pattern of large distances within a certain time period is observed, the estimations can be corrected by adding the mean distance of the relevant data points to the original estimation. This method does not have an effect on the original models, but customer-specific differences are still taken into account.

7.2. Comparison with the previous solutions

Compared to the current dimensioning solution developed in 2014, the new dimensioning tool differs both in principle and in usage. Even though linear regression modeling was utilized in earlier versions, the preceding dimensioning tool was based on predetermined limits due to the lack of data, non-optimal predictor variable choices, and unsuitable tools. The new dimensioning tool uses linear regression extended with an automated subset selection algorithm. This significantly reduces the amount of human effort necessary.

The user interface is completely renewed for the new dimensioning tool. The previous versions have been used with spreadsheet applications that the tools have been developed on. One goal with the new tool was to expand the solution for customer use, which led to a need of a more sophisticated graphical user interface. The solution was a web-based application with appropriate visualization options. Unlike the previous solution, the new graphical user interface also presents the current data on which the predicted resource usage is based on.

Naturally, even if the new tool has a more intuitive user interface, learning is still required from the end user like in most new systems. The new solution offers more information related to the subject. This information is presented with graphs and statistics, which may require some familiarity from the end

users. However, the new tool offers significantly more useful information than the earlier solution.

The developer of the preceding dimensioning tool, Heinonen, mentions some differences between the solutions (2015b). According to Heinonen, the main development problems with the preceding tool were oversimplified capacity modeling, inaccurate modeling for produced load, and inflexible publishing environment. These issues were the starting point for the design of the new dimensioning tool, leading to the creation of the data-driven modeling algorithm running on top of R. The new dimensioning tool meets both the current and the future needs better than the old one.

Heinonen (2015b) emphasizes that the capacity modeling of the preceding dimensioning tool for the network management system was very simple due to the lack of time and indicative data in the implementation phase. The modeling solution fiercely simplifies the load and data in different hardware configurations and does not take all the functionality or customer scenarios into account. Therefore, the capacity modeling of the preceding version can be used only to approximately estimate when the system capacity runs out. As a solution to this problem, the capacity modeling of the new dimensioning tool is entirely based on performance data, in which all measured input variables are included.

Modeling the load produced by the managed network is also problematic in the previous dimensioning solution (Heinonen, 2015b). The modeling is based on theoretical calculations rather than real customer networks which causes inaccurate estimations. Therefore, the amount of resources is kept significantly higher than the estimations required to ensure enough capacity. The new solution produces more accurate estimations and needs less reserve, because data-driven algorithms are used.

The spreadsheet application, on which the preceding dimensioning tool was implemented, restricts the maintenance and expandability of the dimensioning tool (Heinonen, 2015b). Also, implementing a functional and user-friendly user interface proved to be challenging the with the spreadsheet application. Some improvements the users have requested for are not possible without a lot of work. However, the spreadsheet application was a convenient choice, because it was already installed to the workstation of every user. The new dimensioning tool is a web-based application which allows many useful solutions, but is more complex and time consuming to develop.

According to Heinonen, the preceding dimensioning tool supported only some of the possible configurations and enabled answering only some of the questions that arose. Another problem is that due to the shortages and

confidentiality issues, the tool can be distributed only to users inside the company. Unlike its predecessor, the new dimensioning tool has the potential to handle all these challenges.

7.3. Test run with sample data

In this section, the functioning of the new algorithm is introduced and evaluated with a sample of data from multiple comparable simulated network management system environments. The sample data (Appendix 1) has 516 data points and 11 dimensions. As it can be seen from Table 5, where some of the first data points are presented, the data matrix slightly differs from the one introduced in Table 2. Unlike in Table 2, the *timestamp* alone is not the unique key. The unique key is the combination of *source* and *timestamp* variables, where the *source* specifies the simulated network management system from which the data point is from. Nevertheless, because in this case different sources are comparable, all data points can be used for analysis.

Source	*Timestamp*	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
299	12/15/2015 04:00	5579.9	1.24×10^8	1443	1474050	1184743	85922	1022	84	2.97
300	12/15/2015 04:00	5271.6	1.87×10^8	2080	2569286	1762973	89794	1235	73	3.09
301	12/15/2015 04:00	5060.2	2.24×10^8	2647	2986618	2041538	84537	1128	75	3.15
302	12/15/2015 04:00	6161.3	1.65×10^8	1910	1239121	1634375	86564	649	133	2.52
299	12/15/2015 05:00	7172.7	1.50×10^8	1679	2219022	1933074	90263	1321	68	3.17
300	12/15/2015 05:00	4648.9	1.66×10^8	1881	1965165	1319489	88021	1045	84	3.06
301	12/15/2015 05:00	4557.6	8.16×10^7	1029	1391021	2150000	79379	1352	59	3.06
302	12/15/2015 05:00	6385.5	3.03×10^8	3509	2688657	1507277	86338	766	113	2.80

Table 5. The first data points from the sample data.

As in Table 5, the sample data set has one response variable and eight predictor variable candidates. The response variable includes hourly means of the usage of a particular resource. The eight explanatory variables present all the measurable variables that have a possibility to affect the particular response variable. These variables are commonly used for observing the functioning of the network management system. The X_1, X_2, and X_3 variables are total sums in an hour, and the rest are hourly means. Suitable statistics for the data matrix were chosen by person with environment specific knowledge to ensure the

representativeness of the data set. Some statistics for the distributions of the response and predictor variable candidates are presented in Table 6.

	Y	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8
Standard deviation	1801.8	65574348	748.63	1658276	623921	18987	1219.49	30.15	0.49
Mean	3685.5	149329069	1500.6	3077182	1038199	101775	2356.79	55.81	3.14
Median	3711.4	147230671	1411.5	2698453	596019	113471	2756	41	3.23
Min	189.5	0	0	0	0	0	0	0	0
Max	7689.5	338495925	3903	8961610	3211896	120279	5327	177	3.78

Table 6. Some statistics for the variables in the sample.

The models that the algorithm ranked the best are presented in Table 7 in the form introduced in Table 4. From R^2 and adjusted R^2 it can be seen that the differences in explained variance between the best models are very small. This was an expected result: Thepredictor variable candidates partly explain the same variance from different angles. Since multicollinear models were excluded in earlier phase, all models have good variance inflation factor values. The model ranked the best looks like a clear choice to forward with: It has the best value in every statistical indicator.

Resource	*Rank*	*Mallows' C_p*	R^2	*Adjusted* R^2	*F-statistic*
Y	1	45.16	0.76	0.76	411.72
Y	2	54.42	0.76	0.75	323.13
Y	3	56.11	0.76	0.75	268.94
Y	4	57.11	0.75	0.75	268.31
Y	5	60.52	0.75	0.75	266.17
p-value	*Degrees of freedom*	*Residual standard error*	*Intercept*	X_1	X_2
< 0.001	511	885.67	622.94	0	0
< 0.001	510	888.97	635.64	3.14×10^{-5}	-2.10
< 0.001	509	890.62	632.72	0	0.66
< 0.001	509	895.27	551.76	-4.79×10^{-5}	4.07
< 0.001	509	898.89	668.70	7.99×10^{-6}	0
X_3	X_4	X_5	X_6	X_7	X_8
0.0005	0	-0.08	0	41.57	2422.83
0	0	-0.09	0	32.85	2856.88
0.0003	0.0008	-0.03	-0.72	0	1682.28
0.0009	0.001	0	-0.33	18.07	0
0.0002	0.0007	-0.04	-0.72	0	1813.16

Table 7. The output from model generator algorithm with the sample data.

The best models had approximately 75 % of explained variance (Table 7). This means that 25 % of the variation in data is left unexplained by the models. From the sample data set (Appendix 1) it can be seen that even if all predictor variable candidate values are zeros, some differences in the values of the response variable still exists. Because all possible and measurable predictor variable candidates are included, the variation unexplained by the model can be caused by the natural variation in the response variable, by the inaccuracy of the aggregations made in pre-processing phase, or by unmeasurable or inaccessible factors. However, whatever reason the unexplained variation is caused by, the explained variation is quite good.

The best ranked model is visualized in Figure 9. The *Y* axis presents the values for the actual resource usage, while the *X* axis presents the value for the predicted resource usage by the chosen model. The line in the scatter plot shows where the values for both the actual and predicted values are the same. The dots above the line represent cases where the actual value is higher than the predicted one and the dots below the line the other way around. The few outliers of the linear combination of predictor variables that can be seen in Figure 9 are not very distinct and can be left as they are. This type of plot visualizes the portion of unexplained variation in the model well.

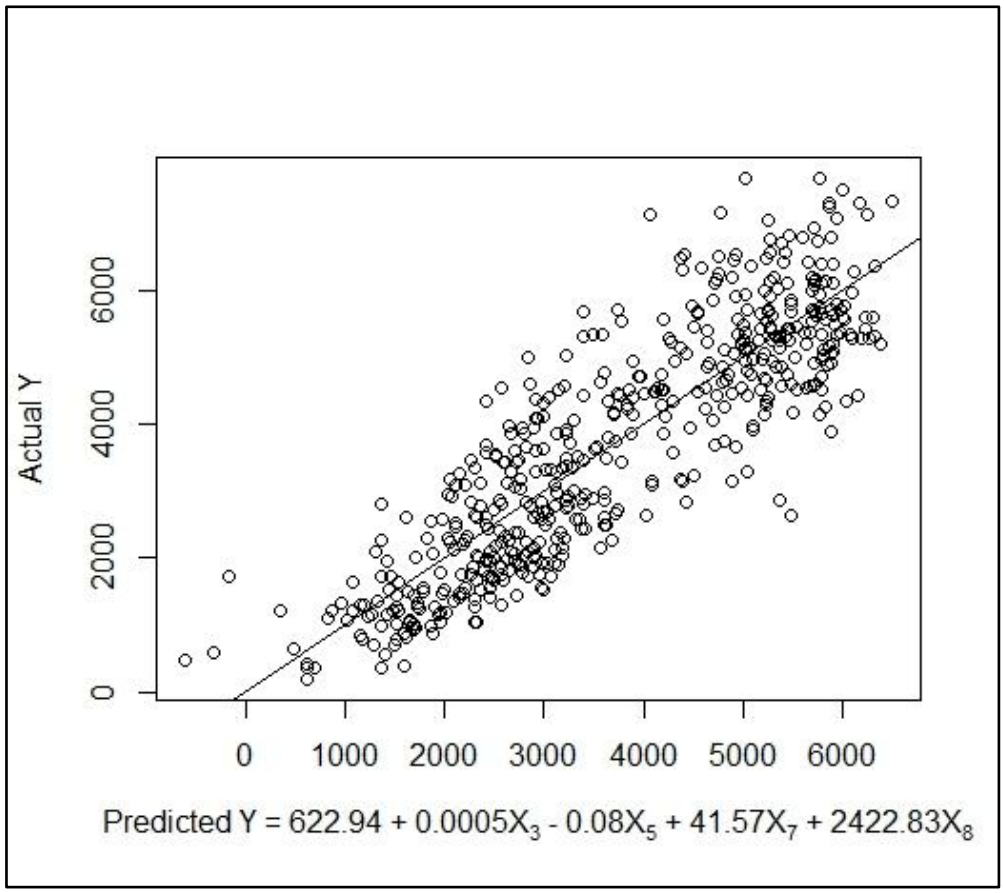

Figure 9. The relationship between the actual and predicted Y.

From Figure 9 it also can be seen that some predictions are negative. These are not realistic in this framework, since negative capacity usage does not exist. At least in this case they do not have a great effect on the regression line, but to ensure the trustworthiness of the result and also to avoid biased predictions later on, the cause for below zero prediction should detected and handled properly. In this case, it seems that the few significantly lower values, still over zero, in X_8 with the negative weight of X_5 cause the situation. This problem will likely be solved once more variation in X_8 is available in model creation, but

another possibility is to remove the data points with outlying values in single predictor variable and then predict only between those limits.

Based on this model, the model generator algorithm works quite well. A linear relationship between the response and predictor variables was detected, and 76 % of the variation can be explained by the resulting model. The issues that were found are likely to be solved once more data with more variation is available, and at same time the model becomes more and more trustable. The model can already be used for estimating the resource usage, as long as the prediction interval is kept in mind. The true value of the algorithm can be determined once all resources are modeled after collecting the suitable data with a lot of variation. However, for this early state in the work, the results look quite promising.

8. Future work

At this point of the process, after many brainstorming sessions and ambitious ideas, a rough division between what must and what can be implemented has been done. The former are already presented in Chapters 6 and 7, while this chapter concentrates on the latter: Ideas, which may or may not be implemented later on in the process, depending on future needs and resources available. Since the work is still in very early state, no detailed plans exist for the proposed features that are presented in this chapter.

The theory behind polynomial and nonlinear relationships, as well as expanding the algorithm functionality to handle them, is discussed in Section 8.1. The decision on whether to implement the algorithm to manage other than linear relationships depends on the features of future data. If it can be seen that the growth in resource usage acts polynomially or nonlinearly, extending the algorithm becomes necessary.

Deeper understanding of the network management system is the aim of the application area introduced in Section 8.2. Discovering and visualizing the strengths of relations between variables is not a part of network management system dimensioning, but utilizes the output of the automated model generation algorithm. The decision of the fate of this feature will be made once the more essential parts of the user interface have been completed.

8.1. Nonlinear relationships

In the network management system framework it is possible, even probable, to find relationships between the response and predictor variables that cannot be modeled with ordinary linear regression models, even though linear regression can in some cases be a relatively good approximation. However, response variables with nonlinear relationships can be important and need to be modeled with appropriate regression technique. Some common nonlinear relationships are, for instance, exponential or logistic. The polynomial regression model, a special case of the general linear regression model, is also introduced in this section.

The ordinary simple or multiple linear regression models include only first-order parameters. The polynomial regression models are considered linear in the parameters, despite of the nature of the response function (Kutner *et al.*, 2005). In polynomial regression models, the parameters can be, for instance, quadratic or cubic, forming regression lines similar to ones presented in Figure 10.

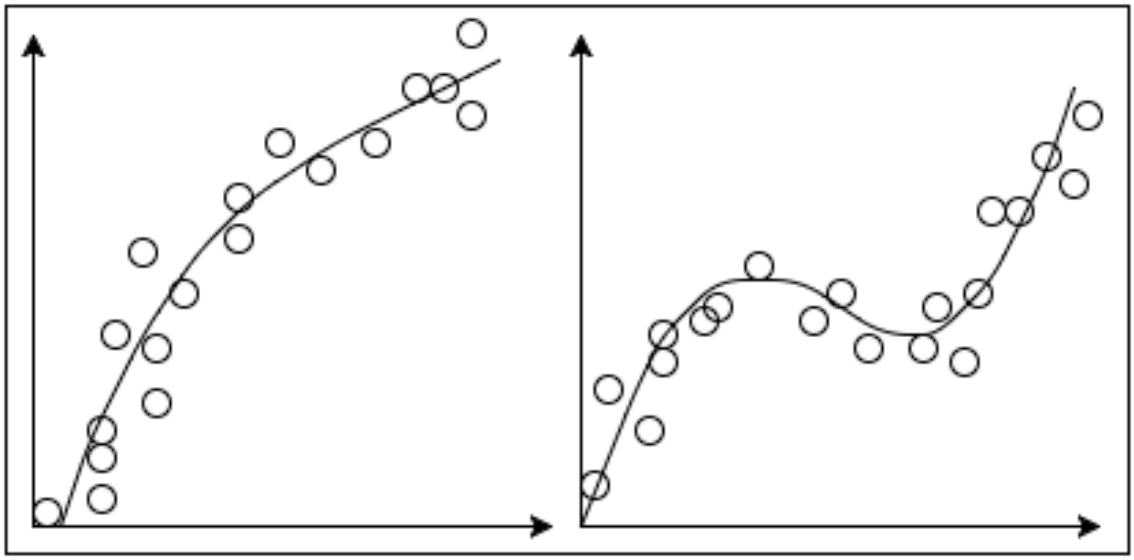

Figure 10. Polynomial regression lines. A quadratic function on the left and a cubic function on the right.

Polynomial regression models can often give good approximations of the training data rather than describe the real data space accurately *(Kutner et al., 2005).* This causes problems when new data do not fit inside the range of the training data. Using polynomial regression models for prediction is still possible if the limits of the response variable are known and the data points are distributed evenly within that range.

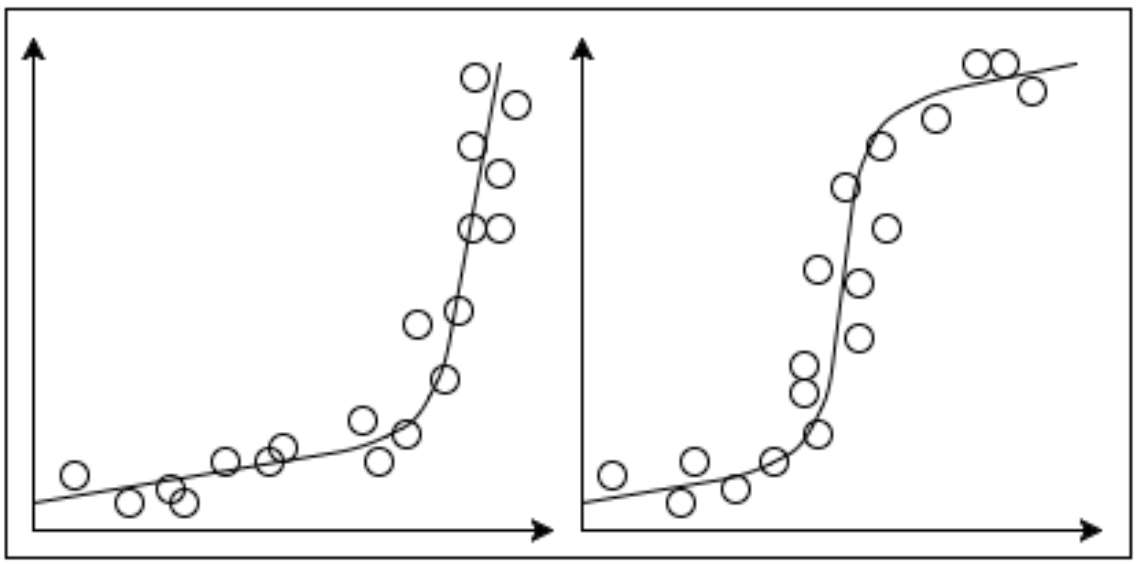

Figure 11. Nonlinear regression lines. An exponential function on the left and a logistic function on the right.

The regression lines of two nonlinear models, the exponential and logistic, are presented in Figure 11. As it can be seen, in the exponential model the growth of the Y value is moderate at the start, but escalates fast once the X

value crosses a certain limit. Many response variables in network management follow the exponential regression curve when all possible values are measured. However, in dimensioning the focal point is in the linearly growing part at the start, which represents the predictable usage before the theoretical maximum point. After the theoretical maximum the resource usage often grows exponentially which is a situation to avoid in network management system dimensioning.

In a logistic model, the growth in the response variable starts moderately, continues rapidly and slows down again, as presented in Figure 11. The logistic regression model is commonly used for estimating binary type variables. Some continuous variables that can naturally be divided to two low-variance clusters are often also treated as binary, and, therefore, the logistic regression model is applied. These variables are quite common in the network management system environment. Some system events are preset to the certain moment of a day, a week or other time period from which it only has values from. The values for those variables from other moments are zero. Since the variation of these variables is not high enough for trustworthy linear regression analysis, logistic regression analysis should be applied instead.

Including polynomial or nonlinear regression models to the subset selection search can be tricky. The computational cost increases tremendously if all different regression models are calculated for all possible subsets to be able to compare the models. A more efficient option would be implementing the actual subset search for only linear models and afterwards calculating if other model fits better. However, this method is not guaranteed to find the best models for the response variables.

8.2. The strength of relations between variables

To fully understand the operation and functionality of Nokia's network management system, a lot is still to be explored. The knowledge on the system helps to predict and prepare for otherwise surprising events and changes. Besides the main purpose of the model generator algorithm, finding models which offer good predictions for resource usage, the same output can be used for getting a better understanding of the system dependencies, for instance, by presenting statistics indicating the strengths of relations between variables.

Firstly, all the predictor variable candidates that affect the certain response variable must be discovered, including those that have been discarded in model creation. This is a challenging task, because the effect of one predictor variable is not always clear and may easily disappear to the variation caused by other factors. Secondly, the strengths of the found relationships must be calculated in

order to find which factors are more important than others. Finally, the relationships and their strengths must be visualized understandably for people unfamiliar with statistical methods. Various different methods can be utilized to complete these tasks (Kutner *et al.*, 2005).

At first, it may seem that the predictor variables in the best model are the variables that affect the response variable, which is not always the correct conclusion. Some predictor variable candidates are discarded in the model creation phase due to multicollinearity or other issues. However, when exploring the relations in the system, the discarded predictor candidate variables can still offer significant information.

One good option for finding the variables that have an effect on the response variable is to utilize the output matrix from the model generator algorithm. The model ranked the best is in the first row, but other models that also exceed the predetermined limits for statistics are included in the output matrix. Therefore, even if a variable with an effect on the response variable is not, for some reason, included in the first model, it is included in one of the models. Thus, the predictor variable candidates with a relationship with the response variable are the variables which have a value other than zero at least in one of the models.

Once the variables with an effect on the response variable are found, the strength of the relation must be measured. The value for the coefficient cannot be used for this purpose, because it depends on the absolute values of the variable (Kutner *et al.*, 2005) in the data and thus is not comparable. An easy solution is to calculate the correlation coefficients between the response variable and all the variables included to any of the models. The correlation coefficients can then be compared (Kutner *et al.*, 2005), and the variables with the most effect on the response variable can be found.

For a wider perspective, the dependencies and their strengths must be visualized. One option is to present the variables in an order from highest to lowest correlation coefficient. However, a problem with this approach is that the order of variables, with close to equal correlation coefficients, may change without real changes in system, depending on the data used for current models. Naturally, this can be confusing to the user. To minimize this effect, the predictor variables from all the models can be classified based on the distribution of the calculated correlation coefficients. For instance, the classes of high, moderate, and low correlation coefficients can be presented to the user together, therefore preventing small changes in correlation coefficients affecting the visualization.

9. Conclusions

In this thesis, an algorithm for modeling capacity usage in network management system environment was designed and implemented. The goal was to utilize the performance data from the system to achieve more flexible and accurate solution for dimensioning. The chosen method was multiple linear regression with an exhaustive branch-and-bound search for subset selection. Also, techniques for external validation, outlier detection, and visualization were specified, even though the available data in this point of the process did not allow thorough evaluation for these methods.

Throughout the work, getting enough representative and varied data was a challenge, because the data collection framework was simultaneously in progress. However, the simultaneous development was also an advantage, because the data collection framework defined the needs for the modeling algorithm and the demands of the modeling algorithm determined some specification in data collection and storing. Due to lack of data, this work formed to be a proof of concept, evaluated with only a fraction of actual data which described the usage of only one resource. However, the resulting algorithm was implemented as a general solution that can be applied to any data in correct format.

Other big challenges during the work were the pre-processing of the data and the evaluation of the resulting models. The pre-processing proofed to be a demanding task, and it was clear that automation is necessary as soon as possible. In evaluating the resulting models, it is clear that external validation must be applied. Determining the goodness of models is not straightforward: It was found out to be characteristic that there is not a one correct answer on how to model the capacity usage, but many close to equal options. Choosing the most trustable model from the offered options may be difficult.

Despite of all the challenges that were faced, a good state with the new network management system dimensioning solution was achieved. The solution reached the point, where the continuation of the work is straightforward, if the quality of data can be ensured. The slow, but inevitable progress finally proofed that dimensioning network management systems by utilizing the performance data is not only possible, but also very educative for understanding the inner dependencies of the system.

References

Alteryx Inc. (2010). *Alteryx designer* (9.5th ed.)

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys, 4*, 40-79.

Berk, K. (1978). Comparing subset regression procedures. *Technometrics, 20*(1), 1-6.

Bollen, K., & Jackman, R. (1985). Regression diagnostics: An expository treatment of outliers and influential cases. *Sociological Methods and Research, 13*, 510-542.

Cook, D. (1977). Detection of influential observations in linear regression. *Technometrics, 19*, 15-18.

Cook, D., & Weisberg, S. (1983). *Residuals and Influence in Regression* (1st ed.). New York, USA: Chapman and Hall/CRC.

Daniel, C., & Wood, F. (1999). *Fitting Equations to Data: Computer Analysis of Multifactor Data* (2nd ed.). New York, USA: John Wiley & Sons, Inc.

Efroymson, M. A. (1960). Multiple regression analysis. *Mathematical Methods for Digital Computers, 1*, 191-203.

Filzmoser, P. (2004). A multivariate outlier detection method. *Proceedings of the Seventh International Conference on Computer Data Analysis and Modeling,* Minsk, Belarus. *1,* 18-22.

Free Software Foundation Inc. (1991). *GNU general public license.*

Freedman, D. (2009). *Statistical Models: Theory and Practice*. New York, USA: Cambridge University Press.

Fujikoshi, Y., Ulyanov, V. V., & Shimizu, R. (2010). *Multivariate Statistics: High-Dimensional and Large-Sample Approximations*. Hoboken, USA: Wiley.

Goldberger, A. (1991). *A Course in Econometrics*. Cambridge, USA: Harvard University Press.

Hand, D., Smyth, P., & Mannila, H. (2001). *Principles of Data Mining*. Cambridge, USA: The MIT Press.

Heinonen, Jyri. (2015a). Personal communication, August 12.

Heinonen, Jyri. (2015b). Personal communication, August 25.

Ihaka, R., & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics, 5*(3), 299-314.

International Organization for Standardization. (1989). *Information Processing Systems - Open Systems Interconnection - Basic Reference Model - Part 4: Management Framework* ISO/IEC 7498-4:1989.

Kariwala, V., Ye, L., & Cao, Y. (2013). Branch and bound method for regression-based controlled variable selection. *Computers & Chemical Engineering, 54*, 1-7.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence,* Montreal, Canada. *2,* 1137-1143.

Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models* (5th ed.). New York, USA: McGraw-Hill Irwin.

Kuusela, Timo. (2015). Personal communication, August 12.

Lempiäinen, J., & Manninen, M. (2002). *Radio Interface System Planning for GSM/GPRS/UMTS*. Norwell, USA: Kluwer Academic Publishers.

Maddala, G. S. (1992). *Introduction to Econometrics*. New York, USA: Macmillan Publishing Company.

Mahalanobis, P. C. (1936). On the generalised distance in statistics. *Proceedings of the National Institute of Sciences of India, 2*(1), 49-55.

Mallows, C. L. (1973). Some comments on cp. *Technometrics, 15*(4), 661-675.

Mantel, N. (1970). Why stepdown procedures in variable selection. *Technometrics, 12*(3), 621-625.

Miller, A. (2002). *Subset Selection in Regression* (2nd ed.) Chapman & Hall/CRC.

Morandat, F., Hill, B., Osvald, L., & Vitek, J. (2012). Evaluating the design of the R language: Objects and functions for data analysis. *Proceedings of the 26th European Conference on Object-Oriented Programming,* Beijing, China. 104-131.

Nokia Solutions and Networks. (2015). *Unpublished intranet content.* Unpublished manuscript.

Oyeyemi, G. M., Bukoye, A., & Akeyede, I. (2015). Comparison of outlier detection procedures in multiple linear regressions. *American Journal of Mathematics and Statistics, 5*(1), 37-41.

R Core Team. (2015). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Rémy, J., & Letamendia, C. (2014). *LTE Standards*. London, UK: Wiley.

Stevens, J. P. (1984). Outliers and influencial data points in regression analysis. *Psychological Bulletin, 95*(2), 334-334.

Tableau software. (2003). *Tableau desktop.*

The R Foundation. (2015). The comprehensive R archive network. Retrieved from https://cran.r-project.org/

Visser, M. D., McMahon, S. M., Merow, C., Dixon, P. M., Record, S., & Jongejans, E. (2015). Speeding up ecological and evolutionary computations in R; essentials of high performance computing for biologists. *PLoS Computational Biology, 11*(3)

Appendix 1: The test run data

source	timestamp	Y	X1	X2	X3	X4	X5	X6	X7	X8
299	2015-12-15 04:00:00	5579.9	1.24E+08	1443	1474050	1184743	85922	1022	84	2.97
300	2015-12-15 04:00:00	5271.6	1.87E+08	2080	2569286	1762973	89794	1235	73	3.09
301	2015-12-15 04:00:00	5060.2	2.24E+08	2647	2986618	2041538	84537	1128	75	3.15
302	2015-12-15 04:00:00	6161.3	1.65E+08	1910	1239121	1634375	86564	649	133	2.52
299	2015-12-15 05:00:00	7172.7	1.5E+08	1679	2219022	1933074	90263	1321	68	3.17
300	2015-12-15 05:00:00	4648.9	1.66E+08	1881	1965165	1319489	88021	1045	84	3.06
301	2015-12-15 05:00:00	4557.6	81680798	1029	1391021	2150000	79379	1352	59	3.06
302	2015-12-15 05:00:00	6385.5	3.03E+08	3509	2688657	1507277	86338	766	113	2.8
299	2015-12-15 06:00:00	6739.9	2.09E+08	2457	2346533	1484394	86029	955	90	3.02
300	2015-12-15 06:00:00	4296.2	1.7E+08	1983	1848942	1485964	86617	933	93	2.88
301	2015-12-15 06:00:00	6130.6	2.08E+08	2349	2108925	1646184	88515	898	99	3.03
302	2015-12-15 06:00:00	5075.7	1.1E+08	1307	1970077	2186243	84432	1507	56	3.08
299	2015-12-15 07:00:00	7305.3	2.2E+08	2530	2788790	1614402	87000	1102	79	3.2
300	2015-12-15 07:00:00	5111.6	1.78E+08	2050	2009584	1631421	86973	980	89	2.84
301	2015-12-15 07:00:00	4733.9	53957484	622	762728	1502217	86748	1226	71	2.96
302	2015-12-15 07:00:00	5356.1	2.47E+08	2883	2702093	1993702	86014	937	92	2.93
299	2015-12-15 08:00:00	5696	1.53E+08	1876	1543039	1785774	81687	823	99	2.68
300	2015-12-15 08:00:00	4492.7	87191971	1035	1469269	3071024	86895	1419	61	3.23
301	2015-12-15 08:00:00	5592	3.03E+08	3427	3129255	1403469	88540	913	97	2.93
302	2015-12-15 08:00:00	5259.1	1.53E+08	1736	2112721	1511176	87873	1217	72	3.2
299	2015-12-15 09:00:00	4958	1.82E+08	2099	1602709	1343704	86596	764	113	2.61
300	2015-12-15 09:00:00	2879.6	1.17E+08	1354	1096349	2192465	86649	810	107	2.85
301	2015-12-15 09:00:00	6498.3	1.95E+08	2244	3284149	1565115	86840	1464	59	3.17
302	2015-12-15 09:00:00	6791.8	2.03E+08	2356	2274366	1669517	86324	965	89	3.12
299	2015-12-15 10:00:00	5637.3	1.77E+08	1999	2714848	1941568	88569	1358	65	3.27
300	2015-12-15 10:00:00	3977.5	1.29E+08	1488	1137837	1361522	86740	765	113	2.63
301	2015-12-15 10:00:00	4573.5	2.03E+08	2460	2508734	1646579	82425	1020	81	2.99
302	2015-12-15 10:00:00	7042.2	1.84E+08	2121	1890899	1543590	89550	892	100	2.86
299	2015-12-15 11:00:00	5849.3	1.28E+08	1481	2184940	1482873	87328	1476	59	3.2
300	2015-12-15 11:00:00	5293.3	2.24E+08	2675	1260175	1522549	83573	471	177	1.88
301	2015-12-15 11:00:00	5360.7	2.35E+08	2699	2950837	1743775	87273	1094	80	3.19
302	2015-12-15 11:00:00	5134.5	1.15E+08	1260	1880068	1924296	91046	1492	61	3.22
299	2015-12-15 12:00:00	6705.7	1.83E+08	2031	3901641	1908606	89981	1921	47	3.42
300	2015-12-15 12:00:00	2632	1.22E+08	1469	982210	1213671	82763	669	124	2.5
301	2015-12-15 12:00:00	4414.8	2.17E+08	2456	2096346	1370275	88236	854	103	2.72
302	2015-12-15 12:00:00	7143.5	1.73E+08	2071	1271606	1958944	83756	614	136	2.58
299	2015-12-15 13:00:00	5459.9	2.4E+08	2735	2202544	1547326	87920	805	109	2.59
300	2015-12-15 13:00:00	6200.8	1.59E+08	1790	2325194	1769469	88932	1299	68	3.32
301	2015-12-15 13:00:00	5943.4	1.71E+08	2120	2330453	1621393	80537	1099	73	3.14
302	2015-12-15 13:00:00	4435.2	1.26E+08	1419	1399486	1703872	90005	986	91	2.92
299	2015-12-15 14:00:00	5622.7	2.27E+08	2624	2188163	2014627	86428	834	104	2.89
300	2015-12-15 14:00:00	4242.7	1.24E+08	1389	1942478	1936100	93005	1399	67	3.27
301	2015-12-15 14:00:00	6781.3	1.83E+08	2216	1923692	1371059	82389	868	95	2.7
302	2015-12-15 14:00:00	4153.4	1.58E+08	1818	2175434	1731114	87045	1197	73	3.16
299	2015-12-15 15:00:00	5364	1.92E+08	2202	1745755	1053987	87176	793	110	2.81
300	2015-12-15 15:00:00	4954.3	1.39E+08	1628	2488847	2127494	85284	1529	56	3.25
301	2015-12-15 15:00:00	4515.3	1.16E+08	1365	1007259	1364628	85124	738	115	2.75
302	2015-12-15 15:00:00	6441.9	2.53E+08	2887	3024655	2261537	87591	1048	84	3
299	2015-12-15 16:00:00	5427	1.53E+08	1816	2363802	1464846	84205	1302	65	3.15
300	2015-12-15 16:00:00	4256.8	2.14E+08	2442	2361318	1377820	87675	967	91	3.09
301	2015-12-15 16:00:00	4644.8	59982446	652	542456	1905665	91998	832	111	2.34
302	2015-12-15 16:00:00	7097.5	2.72E+08	3157	2996878	2119154	86073	949	91	2.95
299	2015-12-15 17:00:00	4562	1.09E+08	1291	1056916	1515014	84070	819	103	2.95
300	2015-12-15 17:00:00	4661.5	1.25E+08	1453	1439875	2300720	86359	991	87	2.89
301	2015-12-15 17:00:00	5115	1.27E+08	1431	2060177	1465215	88577	1440	62	3.26
302	2015-12-15 17:00:00	7322.6	3.38E+08	3903	3712054	1558095	86727	951	91	2.92
299	2015-12-15 18:00:00	5689.9	2.66E+08	3016	2781423	1402517	88104	922	96	2.88
300	2015-12-15 18:00:00	4468.2	64844578	780	855516	2284116	83134	1097	76	3.12

301	2015-12-15 18:00:00	5312.9	1.54E+08	1791	2041024	1276964	85839	1140	75	3.16
302	2015-12-15 18:00:00	5439.6	2.13E+08	2471	2582139	1973517	86332	1045	83	2.98
299	2015-12-15 19:00:00	4920.9	2.05E+08	2465	1907325	1376616	82979	774	107	2.77
300	2015-12-15 19:00:00	5463.5	1.89E+08	2136	1757100	1498056	88519	823	108	2.83
301	2015-12-15 19:00:00	4278.4	91895241	1123	1418300	1569802	81830	1263	65	3.11
302	2015-12-15 19:00:00	6569.6	2.15E+08	2365	3186246	2072835	91052	1347	68	3.2
299	2015-12-15 20:00:00	6287.5	3.23E+08	3668	2637923	1342749	88191	719	123	2.62
300	2015-12-15 20:00:00	4057.4	68087723	837	1150370	2941394	81347	1374	59	3.21
301	2015-12-15 20:00:00	6465.9	1.49E+08	1680	2261291	1933141	88910	1346	66	3.2
302	2015-12-15 20:00:00	4569.9	1.55E+08	1850	2201954	1661685	83572	1190	70	3.16
299	2015-12-15 21:00:00	5619.6	2.19E+08	2472	2719069	1404962	88738	1100	81	3.18
300	2015-12-15 21:00:00	5653.1	1.69E+08	1983	1696682	1340890	86905	856	101	3.01
301	2015-12-15 21:00:00	6192.3	1.86E+08	2139	2219902	1841516	86802	1038	84	2.82
302	2015-12-15 21:00:00	3666.5	1.25E+08	1510	1631826	2100423	82535	1081	76	2.95
299	2015-12-15 22:00:00	4359.3	1.37E+08	1624	1824178	1560469	84183	1123	75	3.11
300	2015-12-15 22:00:00	5711.5	2.49E+08	2803	2774361	1693557	88734	990	90	3.08
301	2015-12-15 22:00:00	6503.8	1.27E+08	1462	1759497	1639119	86546	1203	72	3.06
302	2015-12-15 22:00:00	4694.1	1.87E+08	2184	1908780	1685622	85628	874	98	2.74
299	2015-12-15 23:00:00	4726.1	1.57E+08	1931	1669562	1300654	81121	865	94	2.94
300	2015-12-15 23:00:00	5345.3	56444497	612	1182056	1267808	92230	1931	48	3.3
301	2015-12-15 23:00:00	5369	1.46E+08	1736	2736682	2288511	86707	1576	55	3.26
302	2015-12-15 23:00:00	5615.1	3.31E+08	3750	2651837	1417277	88270	707	125	2.66
299	2015-12-16 00:00:00	5231	1.08E+08	1206	1041344	1150806	89218	863	103	2.72
300	2015-12-16 00:00:00	4639.7	1.88E+08	2078	2750325	2162852	90536	1324	68	3.07
301	2015-12-16 00:00:00	5814	1.41E+08	1770	1472715	1105166	79664	832	96	2.96
302	2015-12-16 00:00:00	5772.8	2.65E+08	3040	3005334	2192865	87008	989	88	3.06
299	2015-12-16 01:00:00	5285.4	2.41E+08	2753	2011033	1348084	87531	730	120	2.81
300	2015-12-16 01:00:00	5917.9	1E+08	1189	2066380	2437557	84489	1738	49	3.39
301	2015-12-16 01:00:00	5484.2	1.92E+08	2348	1738560	1664320	81979	740	111	2.34
302	2015-12-16 01:00:00	5149.1	1.64E+08	1768	2442063	1483948	92688	1381	67	3.32
299	2015-12-16 02:00:00	6552	1.57E+08	1736	2251306	2129924	90487	1297	70	3.2
300	2015-12-16 02:00:00	5575.6	2.67E+08	3092	2665576	1739743	86361	862	100	2.83
301	2015-12-16 02:00:00	5927.1	1.9E+08	2189	2134515	1183527	86960	975	89	2.92
302	2015-12-16 02:00:00	4509.7	84758315	1057	1213838	1727340	80188	1148	70	3.16
299	2015-12-16 03:00:00	5265.7	1.71E+08	1930	1675247	1411739	88804	868	102	2.86
300	2015-12-16 03:00:00	5199.6	1.67E+08	1987	2541743	1830381	84230	1279	66	3.28
301	2015-12-16 03:00:00	4290.7	67468287	740	865196	1536085	91173	1169	78	3.06
302	2015-12-16 03:00:00	7256.4	2.91E+08	3400	3178505	2116237	85712	935	92	2.85
299	2015-12-16 04:00:00	6108.7	1.35E+08	1581	1008397	1130588	85656	638	134	2.55
300	2015-12-16 04:00:00	5712.9	1.83E+08	2029	2128680	2061059	90355	1049	86	2.99
301	2015-12-16 04:00:00	5799.2	2.15E+08	2424	3727885	1844223	88796	1538	58	3.27
302	2015-12-16 04:00:00	5094.9	1.65E+08	2033	1397398	1730138	80945	687	118	2.63
299	2015-12-16 05:00:00	6476.2	1.17E+08	1330	1810515	1345134	88268	1361	65	3.07
300	2015-12-16 05:00:00	6394.6	2.17E+08	2486	1812846	1726626	87361	729	120	2.68
301	2015-12-16 05:00:00	5381	2.17E+08	2515	2141755	1590888	86302	852	101	2.83
302	2015-12-16 05:00:00	4618.8	1.5E+08	1770	2510258	2452872	84629	1418	60	3.34
299	2015-12-16 06:00:00	6796	2.5E+08	2829	3398561	2520089	88279	1201	73	3.11
300	2015-12-16 06:00:00	5493	1.96E+08	2280	1746658	1206992	86368	766	113	2.83
301	2015-12-16 06:00:00	5193.7	1.23E+08	1439	1488532	1839035	85608	1034	83	2.99
302	2015-12-16 06:00:00	4545.6	1.32E+08	1556	1643317	1479174	84826	1056	80	2.98
299	2015-12-16 07:00:00	6539.3	1.61E+08	1782	2075917	2022650	90248	1165	77	2.89
300	2015-12-16 07:00:00	4513.4	1.25E+08	1428	807954	1026901	87656	566	155	2.3
301	2015-12-16 07:00:00	5218.3	3.05E+08	3483	4209147	3211896	87647	1208	73	3.24
302	2015-12-16 07:00:00	5872.8	1.06E+08	1357	1167305	1659148	77968	860	91	2.86
299	2015-12-16 08:00:00	5540.4	1.85E+08	2203	2337030	2409993	84077	1061	79	3.12
300	2015-12-16 08:00:00	5559.6	1.59E+08	1783	1585473	1640265	89437	889	101	2.89
301	2015-12-16 08:00:00	5585.1	1.25E+08	1411	1713303	1890504	88761	1214	73	2.9
302	2015-12-16 08:00:00	4884.8	2.28E+08	2666	2627120	1312566	85606	985	87	3.04
299	2015-12-16 09:00:00	4983.9	90693369	1147	1201900	1335066	79070	1048	75	3.05
300	2015-12-16 09:00:00	5586.6	1.62E+08	1870	1452370	1280931	86723	777	112	2.96
301	2015-12-16 09:00:00	6041.6	2.24E+08	2487	3329427	2084694	90027	1339	67	3.18
302	2015-12-16 09:00:00	5259.8	2.28E+08	2620	2294746	2070037	86839	876	99	2.75

299	2015-12-16 10:00:00	5447.8	1.24E+08	1416	2040830	2836894	87947	1442	61	3.13
300	2015-12-16 10:00:00	3301	1.03E+08	1183	1348282	1930271	87044	1140	76	3.21
301	2015-12-16 10:00:00	7689.5	2.48E+08	2929	2474960	1296972	84614	845	100	2.78
302	2015-12-16 10:00:00	4578.1	2.21E+08	2515	2391920	1377252	87706	951	92	3.02
299	2015-12-16 11:00:00	5421.6	1.7E+08	1986	1867967	1667300	85355	941	91	2.96
300	2015-12-16 11:00:00	4408.5	81544781	951	1544084	2212236	85861	1624	53	3.05
301	2015-12-16 11:00:00	6818.6	2.05E+08	2382	2748502	1401283	85933	1154	74	3.09
302	2015-12-16 11:00:00	4447.7	2.39E+08	2710	2095431	1685937	88189	773	114	2.9
299	2015-12-16 12:00:00	3870.6	55406920	655	1541084	2741788	84801	2353	36	3.43
300	2015-12-16 12:00:00	5709.8	2.7E+08	3196	2500320	1391444	84505	782	108	2.72
301	2015-12-16 12:00:00	5333.7	2.51E+08	2878	1706268	1471870	87374	593	147	2.36
302	2015-12-16 12:00:00	6251.1	1.21E+08	1336	2516747	1871085	90791	1884	48	3.46
299	2015-12-16 13:00:00	6564.5	2.62E+08	2812	3437674	2205345	93152	1223	76	3.14
300	2015-12-16 13:00:00	3928.1	1.76E+08	1988	2156962	1326243	88771	1085	82	3.02
301	2015-12-16 13:00:00	4741.4	91427330	1079	1050503	1334827	84733	974	87	3.17
302	2015-12-16 13:00:00	6089.8	1.73E+08	2229	1630626	1469839	77449	732	106	2.59
299	2015-12-16 14:00:00	4976.3	1.87E+08	2213	1625176	1300056	84425	734	115	2.7
300	2015-12-16 14:00:00	4111.4	1.17E+08	1269	2517656	1730845	92228	1984	46	3.32
301	2015-12-16 14:00:00	4858.5	1.45E+08	1616	2236249	1670727	89517	1383	65	3.2
302	2015-12-16 14:00:00	7335.8	2.5E+08	2966	1879762	1723976	84227	634	133	2.63
299	2015-12-16 15:00:00	5263.1	2E+08	2296	1974094	1899713	86893	860	101	2.82
300	2015-12-16 15:00:00	5435.7	1.44E+08	1740	2028039	1531428	82563	1166	71	3.1
301	2015-12-16 15:00:00	4969.9	1.64E+08	1879	2321768	1672773	87291	1236	71	3.13
302	2015-12-16 15:00:00	5682.8	1.92E+08	2158	1943908	1594608	88975	901	99	2.94
299	2015-12-16 16:00:00	4188.1	1.47E+08	1792	2479999	1487863	82175	1384	59	3.29
300	2015-12-16 16:00:00	6335.7	1.78E+08	2003	2003405	1796625	89019	1000	89	2.73
301	2015-12-16 16:00:00	5977.3	2.19E+08	2538	2046087	1604984	86138	806	107	2.94
302	2015-12-16 16:00:00	4424.2	1.53E+08	1720	1728267	1723686	88857	1005	88	2.99
299	2015-12-16 17:00:00	4870.2	2.19E+08	2489	1634697	1665982	88019	657	134	2.33
300	2015-12-16 17:00:00	6107.8	1.13E+08	1258	1728438	1688227	90313	1374	66	3.28
301	2015-12-16 17:00:00	6947.5	2.2E+08	2586	3504281	1686282	85076	1355	63	3.2
302	2015-12-16 17:00:00	3882.7	1.5E+08	1772	1406676	1520825	84532	794	106	2.95
299	2015-12-16 18:00:00	4869.8	1.86E+08	2122	2971572	1178364	87488	1400	62	3.25
300	2015-12-16 18:00:00	5721	1.45E+08	1699	2242085	1597285	85068	1320	64	3.25
301	2015-12-16 18:00:00	4744.4	95051756	1111	825100	2810677	85555	743	115	2.52
302	2015-12-16 18:00:00	5386.6	2.77E+08	3167	2232665	1923162	87376	705	124	2.6
299	2015-12-16 19:00:00	4407	1.1E+08	1370	1317867	1364654	80112	962	83	2.95
300	2015-12-16 19:00:00	6427.5	1.77E+08	2030	2075592	1419847	87359	1022	85	3.07
301	2015-12-16 19:00:00	5142.5	2.02E+08	2279	2490121	2418809	88630	1093	81	3.08
302	2015-12-16 19:00:00	5048.2	2.08E+08	2375	2378066	1782677	87536	1001	87	2.9
299	2015-12-16 20:00:00	5660	1.65E+08	1838	1459570	1643317	89824	794	113	2.7
300	2015-12-16 20:00:00	5124.8	2.07E+08	2404	2372911	1620959	85901	987	87	3.09
301	2015-12-16 20:00:00	6402.8	2.29E+08	2728	2245493	1557286	83948	823	102	2.83
302	2015-12-16 20:00:00	4531.7	1E+08	1120	2194869	2257835	89452	1960	46	3.28
299	2015-12-16 21:00:00	4594	2.03E+08	2390	2586126	1399293	84834	1082	78	3.13
300	2015-12-16 21:00:00	6175.7	1.15E+08	1345	2228400	2370500	85314	1657	51	3.27
301	2015-12-16 21:00:00	5205.2	2.53E+08	2869	2048078	1521738	88253	714	124	2.64
302	2015-12-16 21:00:00	5072.5	1.31E+08	1493	1410090	1593936	87571	944	93	2.87
299	2015-12-16 22:00:00	5324.7	2.32E+08	2714	1636951	1879143	85710	604	142	2.5
300	2015-12-16 22:00:00	5991.7	2.36E+08	2629	2797849	1792820	89893	1064	84	2.94
301	2015-12-16 22:00:00	5328.5	1.5E+08	1810	1599650	1240437	82624	884	93	2.85
302	2015-12-16 22:00:00	4858.3	77682402	888	2221297	1586053	87480	2501	35	3.58
299	2015-12-16 23:00:00	5171	1.82E+08	2086	2169434	1299726	87319	1040	84	3.25
300	2015-12-16 23:00:00	5683.1	1.46E+08	1585	2256019	2539674	92347	1423	65	3.19
301	2015-12-16 23:00:00	3762.2	1.1E+08	1349	1572655	1036276	83377	1166	71	3.03
302	2015-12-16 23:00:00	6206.5	2.54E+08	3011	2246394	1917592	84451	746	113	2.57
299	2015-12-17 00:00:00	5282.6	58725100	691	1384177	1816896	84986	2003	42	3.39
300	2015-12-17 00:00:00	4348.2	3.09E+08	3527	2651826	1522124	87690	752	117	2.67
301	2015-12-17 00:00:00	5736.9	2.42E+08	2773	3197949	1367352	87359	1153	76	3.14
302	2015-12-17 00:00:00	5952.1	87744893	1073	1026165	2147988	81775	956	86	2.92
299	2015-12-17 01:00:00	6112.2	1.89E+08	2251	2104802	1477558	84140	935	90	3.07
300	2015-12-17 01:00:00	5014.9	2.16E+08	2469	2701251	2230537	89877	1093	82	3.12

301	2015-12-17 01:00:00	4800.2	1.64E+08	1961	1672642	1381150	83427	853	98	2.64
302	2015-12-17 01:00:00	5658.4	1.22E+08	1375	1729614	1607107	88798	1257	71	3.07
299	2015-12-17 02:00:00	6315.9	1.41E+08	1523	1332550	1810618	92799	875	106	2.63
300	2015-12-17 02:00:00	5424.4	2.67E+08	3174	3811349	1401648	84001	1201	70	3.19
301	2015-12-17 02:00:00	4919.2	69558120	818	927850	2015241	85034	1134	75	3.09
302	2015-12-17 02:00:00	5723.1	2.16E+08	2492	2166219	1525243	86528	869	100	2.88
299	2015-12-17 03:00:00	6369	1.39E+08	1641	1840952	1856614	84884	1122	76	3.05
300	2015-12-17 03:00:00	5150.5	1.05E+08	1262	1285833	2992544	83290	1019	82	3
301	2015-12-17 03:00:00	5295.1	3.29E+08	3737	3653706	1249513	87971	978	90	2.96
302	2015-12-17 03:00:00	5340.5	1.21E+08	1379	1468905	1323944	87627	1065	82	3.06
299	2015-12-17 04:00:00	7141	66269111	704	1198998	1782491	94132	1703	55	3.44
300	2015-12-17 04:00:00	4909.5	2.43E+08	2863	2792486	1693483	85048	975	87	2.93
301	2015-12-17 04:00:00	5608.1	3.2E+08	3771	3031448	1288786	84880	804	106	2.76
302	2015-12-17 04:00:00	4148.5	71746765	760	1248432	2224379	94404	1643	57	3.33
299	2015-12-17 05:00:00	7668.7	1.52E+08	1739	2278025	1814456	87691	1310	67	3.19
300	2015-12-17 05:00:00	5396.9	1.7E+08	1950	2025595	1401655	87136	1039	84	2.77
301	2015-12-17 05:00:00	5032.9	2.01E+08	2347	2246515	1795731	85805	957	90	3.06
302	2015-12-17 05:00:00	4148.3	1.75E+08	2028	1711927	1563138	86064	844	102	2.96
299	2015-12-17 06:00:00	6006.1	2.07E+08	2480	2701468	2924078	83669	1089	77	3.07
300	2015-12-17 06:00:00	6148.8	1.96E+08	2250	2295294	1458338	87090	1020	85	3.14
301	2015-12-17 06:00:00	4232.1	1.05E+08	1109	935296	1399928	94455	843	112	2.43
302	2015-12-17 06:00:00	6124.9	1.87E+08	2216	2275183	1927873	85255	1027	83	3
299	2015-12-17 07:00:00	4757.3	1.33E+08	1504	1392958	2069523	88255	926	95	2.93
300	2015-12-17 07:00:00	5317	2E+08	2409	1915441	1862633	82950	795	104	2.63
301	2015-12-17 07:00:00	3946.7	1.3E+08	1347	2837954	1628323	96214	2107	46	3.49
302	2015-12-17 07:00:00	7508.1	2.38E+08	2828	2126682	1292462	84198	752	112	2.73
346	2015-12-14 05:00:00	1114.9	19150885	166	542233	605008	115367	3266	35	3.31
347	2015-12-14 05:00:00	577.1	28719983	259	366289	587937	110888	1414	78	1.98
348	2015-12-14 05:00:00	1432.7	39173777	336	1503333	601015	116589	4474	26	3.59
349	2015-12-14 05:00:00	1202.5	33587355	288	1026145	568534	116623	3563	33	3.55
346	2015-12-14 06:00:00	1068.1	14272341	120	639222	611728	118936	5327	22	3.71
347	2015-12-14 06:00:00	1124.9	36513249	320	947231	599993	114104	2960	39	3.28
348	2015-12-14 06:00:00	1223.1	44934202	398	903415	591000	112900	2270	50	2.91
349	2015-12-14 06:00:00	1154	24912208	211	948132	571187	118067	4494	26	3.65
346	2015-12-14 07:00:00	1371	37659013	326	1027642	603921	115518	3152	37	3.38
347	2015-12-14 07:00:00	655.1	18110088	158	427547	595001	114621	2706	42	3.04
348	2015-12-14 07:00:00	2289.5	43652882	376	1707874	585989	116098	4542	26	3.66
349	2015-12-14 07:00:00	470.6	21076661	186	273589	582150	113315	1471	77	1.98
346	2015-12-14 08:00:00	2618.3	36444481	303	1392118	606536	120279	4594	26	3.78
347	2015-12-14 08:00:00	1655.7	59704078	531	1376635	578464	112437	2593	43	3
348	2015-12-14 08:00:00	374.2	19685324	172	481336	583566	114450	2798	41	3.13
349	2015-12-14 08:00:00	391.3	4931473	46	189259	604215	107206	4114	26	3.57
346	2015-12-14 09:00:00	790.9	8121216	74	334418	564284	109746	4519	24	3.63
347	2015-12-14 09:00:00	1964.4	71327910	624	1621464	590108	114308	2599	44	3.14
348	2015-12-14 09:00:00	1219.5	32516238	276	1142845	595629	117812	4141	28	3.49
349	2015-12-14 09:00:00	371.9	8666636	75	339273	612025	115555	4524	26	3.73
346	2015-12-14 10:00:00	2801.9	63408362	556	1562851	591637	114044	2811	41	3.17
347	2015-12-14 10:00:00	1717.9	76531932	684	992211	579936	111889	1451	77	1.96
348	2015-12-14 10:00:00	2262.7	92597946	818	2013806	586132	113200	2462	46	2.96
349	2015-12-14 10:00:00	1337.7	47649414	420	890100	599535	113451	2119	54	2.9
346	2015-12-14 11:00:00	2938	82878273	701	2575951	602148	118229	3675	32	3.55
347	2015-12-14 11:00:00	4544.9	1.71E+08	1513	3925989	588187	113215	2595	44	3.1
348	2015-12-14 11:00:00	3861.6	1.5E+08	1304	4651863	589865	114684	3567	32	3.43
349	2015-12-14 11:00:00	2302.5	80290465	696	2594242	586697	115360	3727	31	3.45
346	2015-12-14 12:00:00	3280.8	98570597	848	2969266	582340	116239	3501	33	3.41
347	2015-12-14 12:00:00	4109.1	1.72E+08	1491	4681601	588959	115332	3140	37	3.31
348	2015-12-14 12:00:00	2478.4	84174564	734	2402298	604423	114679	3273	35	3.43
349	2015-12-14 12:00:00	3440.3	1.31E+08	1143	3744208	588335	114192	3276	35	3.34
346	2015-12-14 13:00:00	2952.9	79300128	705	2219740	590274	112482	3149	36	3.34
347	2015-12-14 13:00:00	3869.1	1.52E+08	1319	4073987	596018	114959	3089	37	3.34
348	2015-12-14 13:00:00	3851.8	1.34E+08	1180	3979129	584043	115350	3373	34	3.37
349	2015-12-14 13:00:00	2623.9	1.18E+08	1019	3511321	594265	116278	3446	34	3.4

346	2015-12-14 14:00:00	3103.7	75917152	661	2543100	584378	114852	3847	30	3.53
347	2015-12-14 14:00:00	4307.9	1.79E+08	1557	4755639	592201	114961	3054	38	3.26
348	2015-12-14 14:00:00	3942.4	1.47E+08	1295	3989339	585464	113471	3081	37	3.3
349	2015-12-14 14:00:00	2577.9	80671516	683	2463922	602986	118113	3607	33	3.51
346	2015-12-14 15:00:00	3471	1.03E+08	899	3030445	592730	114969	3372	34	3.39
347	2015-12-14 15:00:00	5691.4	2.76E+08	2436	5788374	588222	113154	2376	48	2.98
348	2015-12-14 15:00:00	3992.8	1.48E+08	1306	3931236	583583	113369	3010	38	3.24
349	2015-12-14 15:00:00	2564.9	80649710	677	2573892	602780	119128	3802	31	3.51
346	2015-12-14 16:00:00	4597.9	1.12E+08	987	3802876	579834	113419	3853	29	3.51
347	2015-12-14 16:00:00	5534.3	2.33E+08	2040	6108751	589601	114112	2994	38	3.28
348	2015-12-14 16:00:00	4939.6	2.31E+08	1997	6473859	594883	115705	3242	36	3.34
349	2015-12-14 16:00:00	3660.7	1.48E+08	1270	4242514	595172	116531	3341	35	3.4
346	2015-12-14 17:00:00	4531.9	1.44E+08	1248	4466573	594859	115390	3579	32	3.5
347	2015-12-14 17:00:00	4724.1	2.33E+08	2071	6241510	585349	112549	3014	37	3.29
348	2015-12-14 17:00:00	5315.4	2.12E+08	1821	5788740	594367	116632	3179	37	3.29
349	2015-12-14 17:00:00	3469.4	1.34E+08	1154	4131177	589933	116386	3580	33	3.42
346	2015-12-14 18:00:00	5013.2	2.19E+08	1941	4714654	583560	112887	2429	46	3
347	2015-12-14 18:00:00	3940.4	1.63E+08	1408	4576469	591523	115645	3250	36	3.32
348	2015-12-14 18:00:00	4492.2	2.03E+08	1760	6500004	594762	115301	3693	31	3.5
349	2015-12-14 18:00:00	3718.5	1.39E+08	1184	4836197	595995	117248	4085	29	3.59
346	2015-12-14 19:00:00	4102.1	1.4E+08	1197	4377238	591319	116624	3657	32	3.47
347	2015-12-14 19:00:00	4718.5	2.53E+08	2191	6704881	596020	115425	3060	38	3.28
348	2015-12-14 19:00:00	4473.4	2.36E+08	2074	6344449	587466	113582	3059	37	3.32
349	2015-12-14 19:00:00	3693.9	95830043	833	3202108	584551	115042	3844	30	3.49
346	2015-12-14 20:00:00	4188.9	1.55E+08	1342	5366031	589005	115190	3999	29	3.59
347	2015-12-14 20:00:00	4139.4	2.44E+08	2155	6118333	589445	113057	2839	40	3.17
348	2015-12-14 20:00:00	3527.7	1.9E+08	1655	4123635	590361	114621	2492	46	3.05
349	2015-12-14 20:00:00	4588.7	1.35E+08	1137	5013497	595734	119047	4409	27	3.62
346	2015-12-14 21:00:00	2808.5	1.32E+08	1124	3930530	598251	117169	3497	34	3.4
347	2015-12-14 21:00:00	4422.2	2.51E+08	2204	6262108	593602	114017	2841	40	3.2
348	2015-12-14 21:00:00	4538.9	2.11E+08	1839	6498802	588571	114813	3534	32	3.47
349	2015-12-14 21:00:00	3622.1	1.3E+08	1132	3943064	581902	114995	3483	33	3.4
346	2015-12-14 22:00:00	3382	1.45E+08	1268	4732157	581172	114404	3732	31	3.46
347	2015-12-14 22:00:00	4338	2.41E+08	2118	5844851	590427	113770	2760	41	3.21
348	2015-12-14 22:00:00	4341.2	2.38E+08	2048	7210788	594547	115982	3521	33	3.41
349	2015-12-14 22:00:00	2529.1	1E+08	860	2840204	596721	116548	3303	35	3.4
346	2015-12-14 23:00:00	2216.8	1.12E+08	994	2983229	589073	112970	3001	38	3.23
347	2015-12-14 23:00:00	3285	2.17E+08	1889	5079104	597812	114790	2689	43	3.16
348	2015-12-14 23:00:00	6145.1	2.89E+08	2488	8961610	590607	116059	3602	32	3.47
349	2015-12-14 23:00:00	3304.3	1.05E+08	919	3600655	578943	114727	3918	29	3.51
346	2015-12-15 00:00:00	2630.4	1.08E+08	942	3194987	579913	115072	3392	34	3.37
347	2015-12-15 00:00:00	4513.5	2.57E+08	2233	6993603	588530	114992	3132	37	3.31
348	2015-12-15 00:00:00	3615.3	1.84E+08	1596	4605900	596794	115017	2886	40	3.23
349	2015-12-15 00:00:00	4530.3	1.76E+08	1527	5836912	594259	114947	3822	30	3.53
346	2015-12-15 01:00:00	3096.9	1.33E+08	1158	4297134	592721	114534	3711	31	3.51
347	2015-12-15 01:00:00	4735.1	2.68E+08	2331	7199908	591615	114997	3089	37	3.26
348	2015-12-15 01:00:00	4457.2	2.16E+08	1871	5991142	591435	115322	3202	36	3.36
349	2015-12-15 01:00:00	3125.9	1.07E+08	934	3139692	585641	114915	3362	34	3.4
346	2015-12-15 02:00:00	3341.2	1.33E+08	1165	3326039	592852	114419	2855	40	3.22
347	2015-12-15 02:00:00	3299.2	2.04E+08	1775	4806710	592276	114826	2708	42	3.15
348	2015-12-15 02:00:00	4957.1	2.12E+08	1842	6874012	588884	115323	3732	31	3.5
349	2015-12-15 02:00:00	4770.7	1.72E+08	1510	5617295	590288	115257	3720	31	3.47
346	2015-12-15 03:00:00	2578.1	1.43E+08	1255	4259349	587852	113883	3394	34	3.42
347	2015-12-15 03:00:00	4420.3	2.18E+08	1904	5165386	593254	114332	2713	42	3.11
348	2015-12-15 03:00:00	3810.9	2.05E+08	1777	5844489	593245	115970	3289	35	3.39
349	2015-12-15 03:00:00	3669	1.57E+08	1357	5358064	586373	115690	3948	29	3.54
346	2015-12-15 04:00:00	2879.3	1.96E+08	1728	5669558	585678	113299	3281	35	3.32
347	2015-12-15 04:00:00	5237.9	2.42E+08	2088	7199535	595246	115785	3448	34	3.38
348	2015-12-15 04:00:00	2608.3	1.24E+08	1080	3211018	588607	114811	2973	39	3.28
349	2015-12-15 04:00:00	3352.9	1.62E+08	1398	4547889	588988	116063	3253	36	3.46
346	2015-12-15 05:00:00	2827.9	1.61E+08	1405	4438233	587444	114389	3159	36	3.27
347	2015-12-15 05:00:00	3868.1	2.48E+08	2206	6247099	583990	113527	2832	40	3.23

348	2015-12-15 05:00:00	3867.3	1.55E+08	1344	4899950	593144	115205	3646	32	3.44
349	2015-12-15 05:00:00	4052.2	1.58E+08	1337	5039594	597017	117870	3769	31	3.55
346	2015-12-15 06:00:00	3127.9	1.39E+08	1213	3813373	588473	114761	3144	37	3.32
347	2015-12-15 06:00:00	3497.4	1.94E+08	1656	5748597	595532	116907	3471	34	3.45
348	2015-12-15 06:00:00	3585.9	1.99E+08	1710	6992989	586761	116288	4089	28	3.55
349	2015-12-15 06:00:00	3435.7	1.92E+08	1715	4073041	591620	112033	2375	47	2.98
346	2015-12-15 07:00:00	2872.9	1.19E+08	1025	3685330	592018	116289	3595	32	3.45
347	2015-12-15 07:00:00	3429.7	2.05E+08	1817	4423492	585166	112787	2435	46	3.01
348	2015-12-15 07:00:00	2835.5	2.03E+08	1777	6848711	593037	114438	3854	30	3.54
349	2015-12-15 07:00:00	4444.3	1.96E+08	1675	5669392	594717	117202	3385	35	3.37
346	2015-12-15 08:00:00	3182	1.99E+08	1726	6914546	588624	115556	4006	29	3.56
347	2015-12-15 08:00:00	5335.5	2.28E+08	2009	5485597	588966	113618	2731	42	3.2
348	2015-12-15 08:00:00	2782.3	1.54E+08	1333	3731800	590313	115228	2800	41	3.17
349	2015-12-15 08:00:00	2033.8	1.41E+08	1225	4495184	598698	116214	3670	32	3.43
346	2015-12-15 09:00:00	3171.3	1.13E+08	986	3698681	593407	115074	3751	31	3.51
347	2015-12-15 09:00:00	4477.9	2.66E+08	2324	6913299	591155	114558	2975	39	3.24
348	2015-12-15 09:00:00	2985.7	1.8E+08	1567	4853925	588712	114820	3098	37	3.31
349	2015-12-15 09:00:00	2823.6	1.64E+08	1417	5162095	590198	115859	3643	32	3.48
346	2015-12-15 10:00:00	3590.9	1.21E+08	1054	3479186	591045	114994	3301	35	3.34
347	2015-12-15 10:00:00	3338.5	2.32E+08	2026	5648488	594219	114650	2788	41	3.13
348	2015-12-15 10:00:00	3704.2	2.43E+08	2120	7594661	588898	114736	3582	32	3.48
349	2015-12-15 10:00:00	2227.7	1.27E+08	1094	3905665	588563	116149	3570	33	3.49
346	2015-12-15 11:00:00	3159.9	2E+08	1736	6742257	591223	115215	3884	30	3.57
347	2015-12-15 11:00:00	4347.1	2.11E+08	1853	4271883	593647	114001	2305	49	2.9
348	2015-12-15 11:00:00	3139.9	2.18E+08	1883	6643906	589863	115685	3528	33	3.43
349	2015-12-15 11:00:00	1571.9	94700378	822	2969954	584188	115207	3613	32	3.42
346	2015-12-15 12:00:00	4287.4	2.08E+08	1787	6218779	592143	116542	3480	33	3.45
347	2015-12-15 12:00:00	2965.8	2E+08	1760	5253365	592644	113499	2985	38	3.28
348	2015-12-15 12:00:00	2785.9	1.72E+08	1496	5132964	593440	114860	3431	33	3.37
349	2015-12-15 12:00:00	1959.5	1.44E+08	1251	4022892	581759	115062	3216	36	3.33
346	2015-12-15 13:00:00	2095.7	1.23E+08	1052	3450768	595141	116608	3280	36	3.4
347	2015-12-15 13:00:00	3636.5	2.07E+08	1810	5637768	585414	114111	3115	37	3.28
348	2015-12-15 13:00:00	2670.1	1.88E+08	1617	5965473	591624	116087	3689	31	3.47
349	2015-12-15 13:00:00	3510	1.44E+08	1254	4787991	597407	115051	3818	30	3.54
346	2015-12-15 14:00:00	2423.3	1.14E+08	993	3588755	581574	115014	3614	32	3.43
347	2015-12-15 14:00:00	2753.9	1.81E+08	1572	5096934	593804	115310	3242	36	3.33
348	2015-12-15 14:00:00	3093.5	2.22E+08	1939	6457567	591284	114498	3330	34	3.41
349	2015-12-15 14:00:00	1286.8	85672062	741	2046744	593368	115617	2762	42	3.17
346	2015-12-15 15:00:00	3749	1.72E+08	1485	5686866	594649	115882	3830	30	3.54
347	2015-12-15 15:00:00	2381.5	1.61E+08	1412	4156525	599057	114120	2944	39	3.24
348	2015-12-15 15:00:00	2179.5	1.45E+08	1253	3864412	582382	115773	3084	38	3.28
349	2015-12-15 15:00:00	2017.3	1.25E+08	1095	3482197	590573	114042	3180	36	3.32
346	2015-12-15 16:00:00	3008.3	1.6E+08	1382	5033185	593609	115817	3642	32	3.46
347	2015-12-15 16:00:00	2657.6	1.87E+08	1644	4613307	590513	113715	2806	41	3.17
348	2015-12-15 16:00:00	2858	1.55E+08	1329	4777420	591965	116879	3595	33	3.5
349	2015-12-15 16:00:00	1468	99362442	889	2765160	584934	113274	3111	36	3.3
346	2015-12-15 17:00:00	2236.5	1.25E+08	1074	3648246	594173	115983	3397	34	3.42
347	2015-12-15 17:00:00	3469	2E+08	1767	5406034	579105	113187	3059	37	3.25
348	2015-12-15 17:00:00	2156.4	1.82E+08	1577	5533398	596732	115525	3509	33	3.44
349	2015-12-15 17:00:00	1782.6	96410922	827	2602322	596712	116579	3147	37	3.35
346	2015-12-15 18:00:00	5778.1	2.12E+08	1825	7214846	592635	115984	3953	29	3.56
347	2015-12-15 18:00:00	2258.2	1.87E+08	1599	4381969	593843	116933	2740	43	3.18
348	2015-12-15 18:00:00	1470.9	1.06E+08	939	2787388	587083	113112	2968	38	3.19
349	2015-12-15 18:00:00	1046.6	98300874	882	2805797	585629	111452	3181	35	3.32
346	2015-12-15 19:00:00	3919.5	1.36E+08	1187	4667916	599359	114900	3933	29	3.53
347	2015-12-15 19:00:00	2695.3	1.81E+08	1580	4585449	587586	114783	2902	40	3.27
348	2015-12-15 19:00:00	2576.9	1.86E+08	1628	5127977	592835	114256	3150	36	3.33
349	2015-12-15 19:00:00	979.3	99408232	850	2808658	582604	116951	3304	35	3.32
346	2015-12-15 20:00:00	1549.6	83466965	735	2299766	595211	113560	3129	36	3.26
347	2015-12-15 20:00:00	3434.1	2.02E+08	1764	5734540	597002	114444	3251	35	3.42
348	2015-12-15 20:00:00	2769.8	1.91E+08	1667	5110151	586771	114815	3065	37	3.25
349	2015-12-15 20:00:00	1434.6	1.26E+08	1079	4045324	583104	117164	3749	31	3.49

346	2015-12-15 21:00:00	2708.8	1.37E+08	1178	4459540	592138	116542	3786	31	3.5
347	2015-12-15 21:00:00	3319.6	1.82E+08	1586	4734111	591183	114487	2985	38	3.28
348	2015-12-15 21:00:00	1566.5	1.24E+08	1087	4215317	589210	114378	3878	29	3.5
349	2015-12-15 21:00:00	2494.8	1.6E+08	1394	3781032	590667	114755	2712	42	3.15
346	2015-12-15 22:00:00	2738.8	1.26E+08	1076	3659536	596686	116889	3401	34	3.42
347	2015-12-15 22:00:00	2039.6	1.76E+08	1541	4978413	589374	114275	3231	35	3.32
348	2015-12-15 22:00:00	1915	1.31E+08	1124	4242124	591637	116724	3774	31	3.54
349	2015-12-15 22:00:00	2786.8	1.7E+08	1503	4297856	586799	113095	2860	40	3.2
346	2015-12-15 23:00:00	2324.4	94775164	839	2695655	579151	112962	3213	35	3.36
347	2015-12-15 23:00:00	2595.9	2.16E+08	1894	5055588	591273	114274	2669	43	3.08
348	2015-12-15 23:00:00	1786.1	1.26E+08	1073	4061581	593052	117550	3785	31	3.53
349	2015-12-15 23:00:00	2976.8	1.66E+08	1440	5389176	592574	115228	3742	31	3.51
346	2015-12-16 00:00:00	2024.3	1.06E+08	932	3160169	584637	114106	3391	34	3.36
347	2015-12-16 00:00:00	1874.1	1.8E+08	1558	4158167	597252	115299	2669	43	3.14
348	2015-12-16 00:00:00	2445.7	1.71E+08	1480	5191110	591164	115474	3508	33	3.43
349	2015-12-16 00:00:00	2926.7	1.46E+08	1275	4680554	588700	114727	3671	31	3.49
346	2015-12-16 01:00:00	3543.8	1.22E+08	1056	3501986	595559	115677	3316	35	3.4
347	2015-12-16 01:00:00	2738.5	2.28E+08	1972	6123262	594756	115400	3105	37	3.32
348	2015-12-16 01:00:00	2501.5	1.76E+08	1533	5569813	584027	114487	3633	32	3.44
349	2015-12-16 01:00:00	1490.6	77926957	684	1994939	586566	113928	2917	39	3.22
346	2015-12-16 02:00:00	5723.3	1.49E+08	1285	5539206	589163	115658	4311	27	3.62
347	2015-12-16 02:00:00	2022.4	2E+08	1720	4644237	595624	116221	2700	43	3.2
348	2015-12-16 02:00:00	1963.6	1.5E+08	1303	3839919	592902	114970	2947	39	3.2
349	2015-12-16 02:00:00	1729.9	1.05E+08	937	3166638	580019	111883	3380	33	3.36
346	2015-12-16 03:00:00	5039.7	1.75E+08	1524	5120760	588489	114864	3360	34	3.35
347	2015-12-16 03:00:00	2371.7	1.66E+08	1453	4795189	591151	114497	3300	35	3.36
348	2015-12-16 03:00:00	1874.7	1.46E+08	1265	3571354	581524	115385	2823	41	3.18
349	2015-12-16 03:00:00	2163.7	1.16E+08	1003	3702697	606091	115434	3692	31	3.58
346	2015-12-16 04:00:00	3177	1.09E+08	956	2740145	592569	114364	2866	40	3.24
347	2015-12-16 04:00:00	3226	2.55E+08	2203	7429285	590944	115626	3372	34	3.42
348	2015-12-16 04:00:00	1192.8	98430148	851	2683140	593262	115664	3153	37	3.33
349	2015-12-16 04:00:00	2616.2	1.41E+08	1235	4337430	587325	113905	3512	32	3.38
346	2015-12-16 05:00:00	3500.2	1.22E+08	1051	4560722	600888	116458	4339	27	3.64
347	2015-12-16 05:00:00	2305.1	2.2E+08	1927	5262698	592834	114213	2731	42	3.12
348	2015-12-16 05:00:00	1689.1	1.37E+08	1192	3660303	581083	114666	3071	37	3.28
349	2015-12-16 05:00:00	2077	1.24E+08	1075	3706277	586921	115341	3448	33	3.46
346	2015-12-16 06:00:00	1981.6	81686399	715	2894010	586651	114247	4048	28	3.56
347	2015-12-16 06:00:00	2595.3	2.31E+08	2006	5695093	594246	114976	2839	40	3.18
348	2015-12-16 06:00:00	2500.1	2.02E+08	1744	5652425	590603	115806	3241	36	3.39
349	2015-12-16 06:00:00	1807.4	88867673	780	2948472	586139	113933	3780	30	3.49
346	2015-12-16 07:00:00	2452.5	86817707	750	3067770	593995	115757	4090	28	3.58
347	2015-12-16 07:00:00	2620.9	2.51E+08	2186	6034707	592489	114667	2761	42	3.17
348	2015-12-16 07:00:00	2295.6	1.59E+08	1372	4957999	589596	116240	3614	32	3.45
349	2015-12-16 07:00:00	1960.8	1.06E+08	937	3129524	587040	113338	3340	34	3.39
346	2015-12-16 08:00:00	3624.2	1.45E+08	1230	4433062	610003	117711	3604	33	3.51
347	2015-12-16 08:00:00	2112.5	1.78E+08	1565	4218412	582514	113535	2695	42	3.12
348	2015-12-16 08:00:00	1903.4	1.5E+08	1302	4657153	589194	115255	3577	32	3.46
349	2015-12-16 08:00:00	2393.6	1.31E+08	1148	3881373	585266	113790	3381	34	3.35
346	2015-12-16 09:00:00	3050.1	1.14E+08	992	3862919	590094	114814	3894	29	3.5
347	2015-12-16 09:00:00	1499.5	1.23E+08	1060	2916950	588444	115764	2752	42	3.1
348	2015-12-16 09:00:00	2906.9	2.23E+08	1949	5632834	592672	114608	2890	40	3.24
349	2015-12-16 09:00:00	2895.9	1.43E+08	1244	4777297	590384	115100	3840	30	3.56
346	2015-12-16 10:00:00	2278.4	85158532	754	2701628	568107	112942	3583	32	3.4
347	2015-12-16 10:00:00	2453.7	1.75E+08	1543	5280609	596711	115449	3422	34	3.41
348	2015-12-16 10:00:00	939.2	70987802	604	1835046	614404	117529	3038	39	3.4
349	2015-12-16 10:00:00	2331.8	1.48E+08	1293	3932223	587984	114477	3041	38	3.26
346	2015-12-16 11:00:00	1222	63495094	575	959560	577256	110426	1669	66	2.32
347	2015-12-16 11:00:00	2159.7	1.52E+08	1324	4812338	588472	114943	3635	32	3.43
348	2015-12-16 11:00:00	2627.8	1.87E+08	1597	6310734	594044	117087	3952	30	3.58
349	2015-12-16 11:00:00	1311.4	79861454	700	1669368	601253	114088	2385	48	2.96
346	2015-12-16 12:00:00	2004.2	56118492	476	1925103	590486	117896	4044	29	3.57
347	2015-12-16 12:00:00	2246.7	1.93E+08	1703	4676732	589211	113965	2746	42	3.16

348	2015-12-16 12:00:00	1815.7	1.32E+08	1148	3908050	593568	115400	3404	34	3.38
349	2015-12-16 12:00:00	1300.3	99735543	868	3240084	590914	114903	3733	31	3.52
346	2015-12-16 13:00:00	3318.8	1.3E+08	1118	4235278	591533	115949	3788	31	3.53
347	2015-12-16 13:00:00	1756.9	1.5E+08	1316	3353403	594332	114292	2548	45	3.11
348	2015-12-16 13:00:00	1720.1	1.36E+08	1193	4162177	585172	113770	3489	33	3.37
349	2015-12-16 13:00:00	966.6	65298269	558	1982635	589447	117022	3553	33	3.46
346	2015-12-16 14:00:00	2850.5	1.26E+08	1125	3158483	582306	113241	2808	40	3.2
347	2015-12-16 14:00:00	1915	1.57E+08	1357	4739025	588804	115879	3492	33	3.43
348	2015-12-16 14:00:00	1683.1	1.2E+08	1027	3632396	610521	117187	3537	33	3.49
349	2015-12-16 14:00:00	1267.5	78881825	697	2237963	580379	113173	3211	35	3.25
346	2015-12-16 15:00:00	1653	69155886	592	1944608	591133	116817	3285	36	3.34
347	2015-12-16 15:00:00	1413.4	1.2E+08	1045	3514127	585334	114757	3363	34	3.36
348	2015-12-16 15:00:00	2107.7	1.52E+08	1314	4088575	601579	115465	3112	37	3.37
349	2015-12-16 15:00:00	1865.5	1.42E+08	1245	4204690	580957	113840	3377	34	3.37
346	2015-12-16 16:00:00	2077	78182805	663	2540233	598004	117923	3831	31	3.48
347	2015-12-16 16:00:00	2266.7	2.21E+08	1948	5922000	586669	113547	3040	37	3.27
348	2015-12-16 16:00:00	1987.3	1.28E+08	1106	3646692	593562	115704	3297	35	3.42
349	2015-12-16 16:00:00	860.7	55186715	479	1643075	592481	115212	3430	34	3.41
346	2015-12-16 17:00:00	2047.9	98476013	853	3267779	585427	115447	3831	30	3.45
347	2015-12-16 17:00:00	1745.1	1.47E+08	1291	3217614	596630	114023	2492	46	3.07
348	2015-12-16 17:00:00	2490.1	1.59E+08	1374	4633614	590131	115947	3372	34	3.4
349	2015-12-16 17:00:00	1053.5	77537852	678	2632993	583765	114363	3883	29	3.55
346	2015-12-16 18:00:00	2121.9	95047144	842	2582216	574562	112883	3067	37	3.29
347	2015-12-16 18:00:00	1940.6	1.62E+08	1411	4635060	593798	115123	3285	35	3.36
348	2015-12-16 18:00:00	2121.2	1.67E+08	1440	4867192	594614	115661	3380	34	3.38
349	2015-12-16 18:00:00	1008.4	58491647	503	1667532	600303	116286	3315	35	3.46
346	2015-12-16 19:00:00	1540.4	60274281	530	1637659	588314	113725	3090	37	3.24
347	2015-12-16 19:00:00	1730.1	1.69E+08	1485	4617648	584978	114024	3110	37	3.31
348	2015-12-16 19:00:00	1721.7	1.06E+08	919	3334448	598878	116309	3627	32	3.48
349	2015-12-16 19:00:00	2040.1	1.46E+08	1261	4161756	592495	115728	3300	35	3.38
346	2015-12-16 20:00:00	1220	67016168	591	1734860	578834	113395	2935	39	3.22
347	2015-12-16 20:00:00	3143.7	2.65E+08	2260	8304208	598388	117177	3674	32	3.49
348	2015-12-16 20:00:00	1154.9	83159765	736	1909144	581889	112989	2594	44	3.03
349	2015-12-16 20:00:00	884.9	67531370	609	1803788	581160	110889	2962	37	3.29
346	2015-12-16 21:00:00	2843.4	1.16E+08	1000	4056387	597383	116376	4056	29	3.6
347	2015-12-16 21:00:00	1465.5	1.63E+08	1431	3511174	590230	114044	2454	46	3.02
348	2015-12-16 21:00:00	1442.1	95636981	850	2345943	579659	112514	2760	41	3.09
349	2015-12-16 21:00:00	1917.3	1.07E+08	915	3838496	596389	117287	4195	28	3.6
346	2015-12-16 22:00:00	2086.5	95256182	829	1979895	596459	114905	2388	48	2.97
347	2015-12-16 22:00:00	1517.9	1.37E+08	1197	4271452	590778	114597	3568	32	3.45
348	2015-12-16 22:00:00	1378.3	86553670	743	2955103	591513	116492	3977	29	3.56
349	2015-12-16 22:00:00	838.9	63778030	565	1369550	577661	112881	2424	47	2.98
346	2015-12-16 23:00:00	2235.5	1.16E+08	1015	2927156	589030	114763	2884	40	3.21
347	2015-12-16 23:00:00	1068.3	76665306	667	1945527	596354	114940	2917	39	3.27
348	2015-12-16 23:00:00	1277.3	66932994	579	2097081	589727	115601	3622	32	3.49
349	2015-12-16 23:00:00	1710	1.02E+08	886	3344236	589336	114914	3775	30	3.48
346	2015-12-17 00:00:00	1299	42522438	370	1048984	590053	114926	2835	41	3.22
347	2015-12-17 00:00:00	1348.7	1.23E+08	1085	2909798	592148	113645	2682	42	3.16
348	2015-12-17 00:00:00	1058.9	75626383	673	1957357	577193	112372	2908	39	3.18
349	2015-12-17 00:00:00	1749.6	1.2E+08	1019	4397861	599185	118196	4316	27	3.62
346	2015-12-17 01:00:00	1718.4	68311121	600	1777121	587147	113852	2962	38	3.21
347	2015-12-17 01:00:00	1671.2	1.11E+08	941	3288114	597844	117645	3494	34	3.45
348	2015-12-17 01:00:00	1746.3	1.07E+08	935	3372690	590889	114223	3607	32	3.45
349	2015-12-17 01:00:00	958.6	76082625	671	1876075	584829	113387	2796	41	3.21
346	2015-12-17 02:00:00	3095	71259328	628	2373388	570065	113470	3779	30	3.48
347	2015-12-17 02:00:00	1910	1.73E+08	1514	4375728	588890	114061	2890	39	3.2
348	2015-12-17 02:00:00	994.2	67118075	571	1649716	605546	117545	2889	41	3.27
349	2015-12-17 02:00:00	1031.3	50830993	434	1915168	595334	117122	4413	27	3.69
346	2015-12-17 03:00:00	4377.8	1.1E+08	948	4033247	595203	116245	4254	27	3.61
347	2015-12-17 03:00:00	1172.4	1.05E+08	912	2689497	585123	114717	2949	39	3.23
348	2015-12-17 03:00:00	690.3	46157516	400	1377477	592429	115394	3444	34	3.34
349	2015-12-17 03:00:00	1324.1	1.01E+08	887	2213779	589220	113773	2496	46	3.09

346	2015-12-17 04:00:00	4086.8	1.12E+08	954	4187654	596039	117413	4390	27	3.63
347	2015-12-17 04:00:00	1541.5	1.11E+08	977	2577045	593629	113564	2638	43	3.17
348	2015-12-17 04:00:00	1012.9	92195921	807	2081433	582542	114245	2579	44	3.08
349	2015-12-17 04:00:00	566.3	46736872	409	1467868	585332	114271	3589	32	3.37
346	2015-12-17 05:00:00	3073.8	1.32E+08	1159	3766645	595031	113870	3250	35	3.36
347	2015-12-17 05:00:00	1403.8	1.14E+08	998	2890545	587350	114311	2896	39	3.26
348	2015-12-17 05:00:00	712	66902961	575	1952506	590070	116353	3396	34	3.34
349	2015-12-17 05:00:00	810.7	48935221	415	1704304	590385	117916	4107	29	3.58
346	2015-12-17 06:00:00	797.3	38863883	340	1236567	581276	114306	3637	31	3.34
347	2015-12-17 06:00:00	1253.6	1.27E+08	1113	2670455	589216	113844	2399	47	2.99
348	2015-12-17 06:00:00	1683.5	97057956	841	3105724	593183	115408	3693	31	3.53
349	2015-12-17 06:00:00	1546.2	99265382	853	3301254	595652	116372	3870	30	3.53
346	2015-12-17 07:00:00	1720.4	70421655	605	1893455	593200	116399	3130	37	3.25
347	2015-12-17 07:00:00	1532.8	1.06E+08	915	3302470	596755	115929	3609	32	3.44
348	2015-12-17 07:00:00	1183.8	97863345	857	2493860	586969	114193	2910	39	3.26
349	2015-12-17 07:00:00	1264.4	87536405	770	2624215	582965	113684	3408	33	3.46
346	2015-12-11 11:00:00	347.6	0	0	0	0	0	0	0	0
347	2015-12-11 11:00:00	189.5	0	0	0	0	0	0	0	0
348	2015-12-11 11:00:00	191.6	0	0	0	0	0	0	0	0
349	2015-12-11 11:00:00	194.7	0	0	0	0	0	0	0	0
346	2015-12-11 12:00:00	423.5	0	0	0	0	0	0	0	0
347	2015-12-11 12:00:00	199.8	0	0	0	0	0	0	0	0
348	2015-12-11 12:00:00	202.6	0	0	0	0	0	0	0	0
349	2015-12-11 12:00:00	204.7	0	0	0	0	0	0	0	0