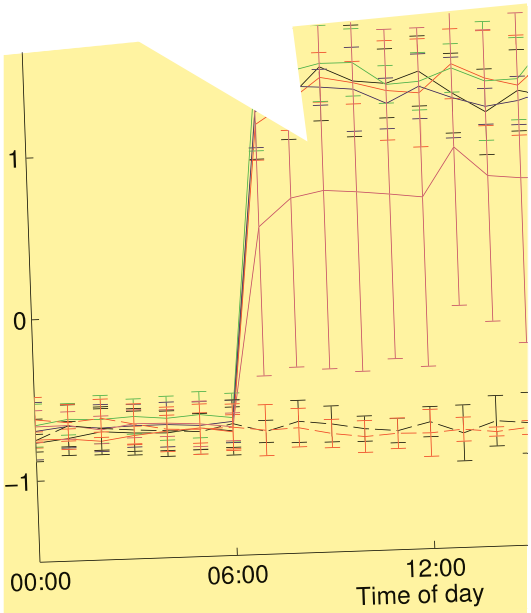
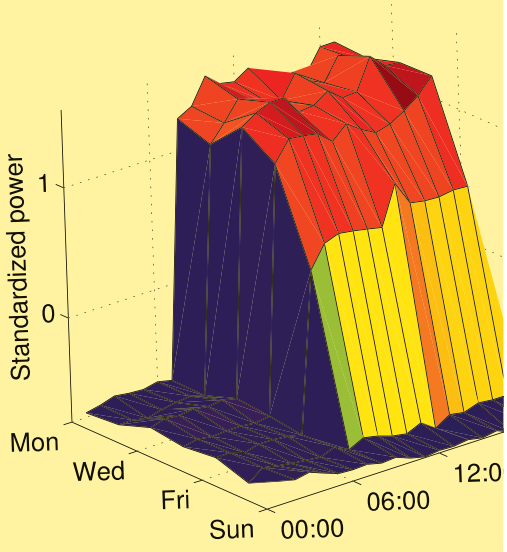
FAULT DETECTION WITH HOURLY DISTRICT DATA

FJÄRRSYN

RAPPORT 2013:27



Average standardized power, P_s, for the off
– 3.3. This figure is generated from one ye
at the outdoor temperature is less than -5°



6. Surface plot of the average standardize
s that illustrated in Figure 3.5.

FAULT DETECTION WITH HOURLY DISTRICT ENERGY DATA

PROBABILISTIC METHODS AND HEURISTICS FOR AUTOMATED DETECTION AND RANKING OF ANOMALIES

FREDRIK SANDIN JONAS GUSTAFSSON JERKER DELSING

ISBN 978-91-7381-125-5

FÖRORD

Inom en relativt snar framtid kommer det bli krav på att fjärrvärme och fjärrkyla ska debiteras efter faktisk användning. Det ställer i sin tur krav på att energileverantörerna kan säkerställa att det som debiteras bygger på korrekta värden. För att detta inte ska bli en onödigt stor administrativ börda för företagen är det viktigt att själva valideringen av mätdata kan ske automatiskt.

Denna rapport beskriver olika statistiska metoder för automatisk detektering av potentiella mätfel. Den information som används är vanligtvis tillgänglig i moderna fjärrvärmecentraler. För att ta fram och utvärdera metoderna har projektet använt historiska mätdata från närmare tusen fjärrvärmecentraler.

Arbetet har utförts av Fredrik Sandin, Jonas Gustafsson och Jerker Delsing på Eislab, Luleå tekniska universitet. Till projektet har en referensgrupp varit knuten. Referensgruppen har bestått av Jan Berglund Mälarenergi, Per Malmberg Processvision, Anders Ricknell Processvision, Roland Lundberg One Nordic, Lars-Ove Ivarsson Vattenfall, Robert Eklund Södertörns Fjärrvärme, Torsten Olsson Göteborg Energi, Lars Lindström Powel och Martin Brage Jönköping Energi.

I projektet har också ingått att utveckla metoder för att kommersialisera forskningsresultat, vilket Jonas Gustafsson redovisat i en separat rapport. Jerker Delsing har deltagit som svensk representant i det internationella standardiseringsarbetet kring energimätare vilket redovisats separat.

Projektet ingår i forskningsprogrammet Fjärrsyn som finansieras gemensamt av Energimyndigheten och fjärrvärmebranschen. Fjärrsyn ska stärka möjligheterna för fjärrvärme och fjärrkyla genom ökad kunskap om fjärrvärmens roll i klimatarbetet och för det hållbara samhället till exempel genom att bana väg för affärsmässiga lösningar och framtidens teknik.

Bo Johansson
Ordförande i Svensk Fjärrvärmes teknikråd

SAMMANFATTNING

Validering av mätdata

Det här projektet har genomförts eftersom fjärrvärmeföretagen anser att det är svårt och kostsamt att detektera fel i stora energisystem. Fel som förblir oupptäckta kan vara kostsamma och branschen förlorar trovärdighet när kunder upptäcker fel och får felaktiga räkningar. Fel är vanliga eftersom systemen innehåller ett stort antal instrumenterade centraler för fjärrvärme och fjärrkyla. Dessutom är den konventionella instrumenteringen konstruerad för fakturering och låg inköpskostnad, inte för automatiserad detektering av fel. Stora variationer i byggnaders dynamik och delsystem, mänskligt beteende och utomhusklimat gör det svårt att modellera och analysera systemen. Konventionella metoder för detektering av fel används därför inte, men enkla gränsvärdesmetoder är vanligt förekommande och kan resultera i ett stort antal falsklarm som är kostsamma att analysera och hantera. Det finns ett växande intresse inom branschen för att utveckla tjänster och funktioner som är baserade på energimätdata med hög tidsupplösning. Regler för energimätning förväntas också bli mer krävande i framtiden, vilket driver den tekniska standarden mot högre tidsupplösning i energimätdata. Denna trend leder till stora dataströmmar på systemnivå, vilket är utmanande när man skall förvissa sig om att data är korrekt. Därför behövs nya effektiva metoder för detektering av fel.

Den här projektrapporten beskriver en rad probabilistiska metoder för automatiserad detektering av avvikelser som är användbara för att kunna identifiera potentiella fel i storskaliga fjärrvärmesystem. Metoderna är förenliga med den information som finns tillgänglig i moderna insamlingssystem för energimätdata. Vi fokuserar på metoder som kan tillämpas automatiskt, med ett minimum av mänsklig assistans för att möjliggöra kostnadseffektiv analys av data. Med hjälp av dessa metoder behöver operatörerna inte spendera tid på ad-hoc tester eller visuell inspektion av grafer för att upptäcka avvikelser i data. Istället kan operatörerna fokusera på att analysera de centraler som identifieras som mest avvikande. Dygns- och veckocykler i effekten modelleras genom att automatiskt gruppera veckans timmar beroende på om effektbehovet är högt eller lågt. Alternativt så kan endast veckocykler modelleras genom att gruppera veckodagar med liknande effektbehov. Robust regression används för att modellera variabelsamband med historiska data. En robust metod för detektering av uteliggare används för att bestämma om variabler avviker från de väntevärden som definieras av regressionsmodellerna. Robusta statistiska metoder används för att bestämma hur mycket uteliggare avviker från de förväntade värdena, så att onormalt avvikande mätvärden kan identifieras och rangordnas automatiskt. Regressionsmodellerna kan också användas för skattning av saknade energimätdata, vilket är ett vanligt problem som inte alltid hanteras på ett noggrant sätt. Vi presenterar också metoder för detektering av drivande signaler, vilket är en typ av fel som kan vara kostsam och i annat fall svår att upptäcka, samt för detektering av bristande

precision i mätdata, vilket till exempel kan vara ett resultat av överdimensionerade flödesventiler, felaktig konfiguration och brus. Förutom att upptäcka fel så kan de föreslagna metoderna även vara användbara för underhållsplanering, eftersom centraler som beter sig på ett förväntat sätt kan ges lägre prioritet jämfört med centraler med onormalt beteende.

Vi studerar metoderna med timvärden från en population om cirka ett tusen fjärrvärmecentraler. Exempel-kod för viktiga metodfunktioner tillhandahålls. Med metodernas hjälp hittar vi onormala data för ungefär 5% av centralerna och bland dessa identifieras dokumenterade fel, okända fel och onormala egenskaper. Saknaden av ett väldefinierat datamängd gör utveckling och utvärdering av metoder för detektering av fel utmanande, och det faktum att historiska energimätdata kan innehålla felaktiga data ignoreras ofta i litteraturen. De föreslagna metoderna måste implementeras i ett fullskaligt fjärrvärmesystem under överinseende av erfarna operatörer innan möjligheterna att detektera fel på ett kostnadseffektivt sätt kan utvärderas. Vi är dock övertygade om att de föreslagna metoderna kan implementeras i nuvarande system för insamling och analys av energimätdata och att de ger väsentliga fördelar jämfört med de metoder som vanligtvis används idag.

SUMMARY

This project is motivated by the difficulties experienced by district energy utilities to detect faults in large-scale district energy systems. Faults that remain undetected can be costly and the industry loose credibility when customers detect faults and receive incorrect bills. Faults are common in district energy systems due to the high number of substations and instrumentation components. Also, the standard energy-metering instrumentation is designed for low cost and billing, not for automated fault detection. Large variations in building dynamics, building subsystems, human behaviour and the environment make the system complex to model and analyse. Therefore, conventional methods for fault detection are not applicable and the use of ad hoc methods for fault detection often result in numerous false alarms that are costly to analyse and manage. There is a growing interest among the utilities to develop services and functions that are based on data with high temporal resolution. Energy metering regulations are also expected to become more demanding in the future, which drives technology standards towards high-resolution data. This trend results in high rates of streaming data at the management level, which is more challenging to validate. Therefore, more efficient methods for fault detection are needed.

This project deals with probabilistic methods for automated anomaly detection that are useful for the identification of faults in large-scale district energy systems. These methods are compatible with the information that is available in modern energy meter data management systems. We focus on methods and heuristics that can be applied automatically with a minimum of human assistance to enable cost-efficient analysis of data. With these methods, operators do not have to rely on ad hoc tests or manual inspection of graphs to detect anomalies in the data. Instead, operators can focus on the analysis of a subset of substations that are identified as abnormal. Intraday and intraweek variations in the thermal load are accounted for by automatically grouping hours of the week with similar thermal load characteristics. Alternatively, intraweek cycles can be accounted for by grouping days of the week with similar characteristics. Robust regression is used to model variable relationships with historical data. A robust outlier detection method is used to determine if variables deviate from the expectation defined by a regression model. Robust statistical methods are used to score outliers, so that outstanding substations can be identified automatically with a ranking procedure. The regression models can also be used for imputation of missing energy metering data, which is a common problem that is not always solved in an accurate way. We also present methods for the detection of long-term drift, which can be costly and otherwise difficult to detect, and the detection of poor precision in measurement data, which for example can result from oversized flow valves, misconfiguration and noise. In addition to fault detection, the proposed methods can be useful also for maintenance scheduling because substations that behave in a way that is consistent with the historical record can be given a lower maintenance priority compared to substations with abnormal behaviour.

The proposed methods are studied using hourly data from a population of about one thousand district heating substations. Sample code of key functions is provided. We find that substations with documented faults, unknown faults and abnormal characteristics can be identified in about 5% of the substations. The lack of a well-defined dataset makes the development and evaluation of methods for fault detection challenging, and the fact that historical energy metering data includes abnormal data is often ignored in the literature. The proposed methods need to be implemented in a full-scale district energy management system under the supervision of experienced operators before the effects on the fault detection rate and cost efficiency can be properly evaluated. However, we are convinced that the proposed algorithms can be implemented in present data management systems and that they offer significant advantages over the methods that are commonly used today.

NOMENCLATURE

Common functions and variables

Symbol	Description
t	Time.
T_{ps}	Primary supply temperature.
T_{pr}	Primary return temperature.
ΔT	Primary temperature difference, $T_{ps} - T_{pr}$, the traditional “delta-T”.
ΔT_{ps}	Difference of primary supply temperatures of two different substations, which are matched by time-series correlation analysis.
m, V	Mass flow, flow. Derived from the volume calculated by the energy meter.
P	Thermal power. Derived from the energy calculated by the energy meter.
T_{out}	Outdoor temperature, typically estimated from meteorological data.
$\mu, E[x]$	Mean value, expectation value.
σ	Standard deviation, square root of variance, $\sqrt{E[(x-\mu)^2]}$.
γ	Skewness, $E[\ x-\mu\ ^3/\sigma^3]$.
κ	Kurtosis, $E\ \ x-\mu\ ^4/E[(x-\mu)^2]^2$.
BC	Bimodality coefficient, $(\gamma^2+1)/\kappa$ (heuristic).
CS	Maximum cumulative sum, $\max(S^+, S^-)$.

Terms and abbreviations

Term	Description
Bimodal (PDF)	A probability density function with two distinct maxima.
De-trending	Method to reduce the significance of a trend in time series data. The trend is a gradual change of some property of the time series over the whole interval under investigation, such as an annual cycle in the outdoor temperature.
Diagnosis	Identification of the nature and cause of something, for example outliers in a dataset.
GESD	Generalized Extreme Studentized Deviate (test for outliers).
Heuristic	Experience-based technique for problem solving (rule of thumb, educated guess, intuitive judgment, common sense).
K-means	A cluster analysis method that partitions observations into clusters in which each observation belongs to the cluster with the nearest centre point.
Modified Z score	An outlier score that is robust (less sensitive to outliers in the data).
Multimodal (PDF)	A PDF with more than one local maximum (see also bimodal).
Outlier	An observation that appears to deviate markedly from other members of the sample in which it occurs.
PDF	Probability density function (of a random variable).
Power	Thermal power. The finite difference (derivative) of the accumulative energy calculated by an energy meter.
Robust	A statistic or estimator capable of coping with outliers (for example, the median is robust but the mean is not robust).

Statistic	A measure of some attribute of a sample, which is calculated by applying a function or statistical algorithm to the values of the items in the sample (for example, mean and standard deviation).
Z score	A statistical score that indicates by how many standard deviations an observation is above or below the expectation value.

TABLE OF CONTENTS

1 INTRODUCTION

District heating systems are commonly used in several countries world wide for space- and water heating in residential and commercial buildings, and for industrial heating purposes. For example, about 90% of the apartment buildings and more than 50% of the buildings in Sweden are heated in that way (Svensk Fjärrvärme, 2013). Similarly, district cooling is used for space cooling and industrial cooling purposes, but it is less common than district heating. Carbon emissions, pollution and the consumption of primary resources can be reduced with the use of district heating plants compared to local, decentralized heat production. In particular the use of surplus heat from industries, waste incineration and combined heat and power production are examples of strategies to minimize the environmental effects and use of primary resources. For a general introduction to district heating, see Frederiksen & Werner (2001).

A district energy system can include tens of thousands of buildings, which comprise the instrumentation needed for energy metering and billing. District energy utilities can be overwhelmed with the quantity of energy metering data, in particular when modern energy meters and data management systems with a temporal resolution of days or hours are used. The standard instrumentation is designed for local control, billing and low cost, not for automated fault detection and diagnosis. Therefore, it is difficult to detect faults in the instrumentation and installation. In addition, some components needed in the energy meter instrumentation can be error-prone, such as mechanical flow meters and temperature sensors.

Instrumentation faults leading to an offset or outliers in the hourly or daily thermal power consumption are not uncommon. We have identified faults resulting in outliers in the hourly average power that deviates by several orders of magnitude from the expectation, and offsets of 100% or more. Such faults lead to incorrect billing if they are not detected, which has consequences for customer relations, trust and costs. It also efficiently hinders the development of information services and business based on energy metering data because of the risks associated with exposing incorrect information. Other faults, in particular affecting the flow meter or amplifiers / voltage references can cause long-term drift of signals and the thermal power. The long time-scale associated with such faults can lead to substantial economic loss. Long-term drift can remain undetected for several years and becomes apparent only when the instrumentation is replaced or serviced, resulting in a sudden change of the power. There are documented cases when the thermal power has increased by 100% after service of the instrumentation, which means that the cost of such faults can be substantial. Also, the slow and gradual change associated with long-term drift makes these faults difficult to detect if simplistic methods are used.

In general, district energy utilities find it difficult to detect faults because of the system size, the deluge of data and insufficient fault-detection functionality in the instrumentation and data management systems. Alarm and warning systems that

implement basic fault-detection tests are commonly used and often result in a deluge of false alarms that are difficult and expensive to manage (Reference group). The commonly used methods are simplistic and depend on the expertise of the operators. The problem to detect potential faults is important for several reasons; Faults that affect billing and customer information services need to be avoided; Energy market regulations tend to become more demanding with time in terms of accuracy, system efficiency and environmental effects; There is an increasing interest among the utilities to exploit the energy metering data for system optimization and development of new information services, in order to stay competitive on the energy market and develop good customer relations. These are key factors that motivate the project that is summarized in this report, which deals with probabilistic methods for automated detection and ranking of anomalies / faults in district energy data. Before describing these methods we introduce central concepts related to district-heating substations, hourly energy metering data, common faults, related work and the aims and scope of the project.

1.1 District heating substations

A district energy system can include tens of thousands of district heating *substations*, which transfer the heat from the distribution network to a building (or any other process that requires heat), see Figure 1.1.

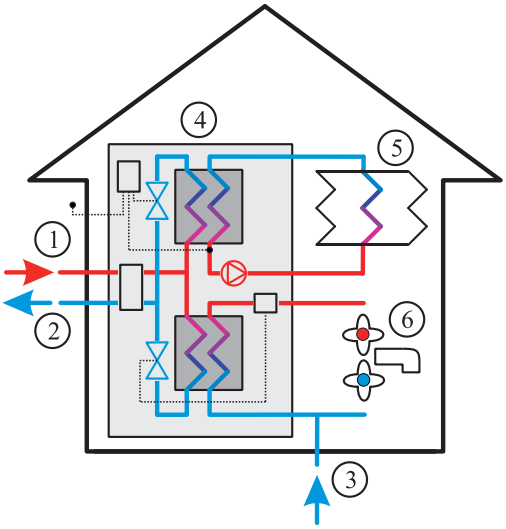


Figure 1.1. Schematic illustration of a building with a district-heating substation. Indicated in the figure are the primary water supply (1); primary return water (2); tap water supply (3); district heating substation including heat exchangers, electronic energy meter and control system with related sensors (4); heating system (5); and tap water (6).

A district-heating substation comprises an electronic energy meter, which calculates the thermal power received from the distribution network. The energy meter includes a flow meter and two temperature sensors, one for the primary supply temperature

and one for the primary return temperature. The thermal energy is calculated from the flow and primary temperature difference

$$E = \int c\, m\, t' \left(T_{ps}\, t' - T_{pr}\, t' \right) dt', \tag{1}$$

where c is the specific heat of the liquid in the distribution network (typically water), m is the mass flow, T_{ps} is the primary supply temperature and T_{pr} is the primary return temperature. The substation also includes heat exchangers and a control system with related sensors and actuators.

The standard instrumentation in a district-energy substation is designed so that there is a clear distinction between the energy meter on the supply side of the heat exchanger(s), which is the property of the energy utility, and the control system of the substation, which belongs to the building. This setup is not ideal for fault detection because the control system has access to valuable information that is not accessible via the energy meter, such as the local outdoor temperature and control set points. The variables that are commonly available in modern energy meter data management systems are the accumulated energy and power, the primary supply and return temperature, and the primary accumulated volume and flow. The temporal resolution ranges from minutes to months, with daily averages being the norm and hourly values becoming increasingly common. The local outdoor temperature, which plays a major role for space heating, is unknown to the energy meter and data management system. Therefore, the outdoor temperature is estimated from meteorological data or proprietary temperature sensors that are deployed within the geographical area of the network.

A district-heating substation is a challenging process for fault detection and diagnosis because it depends on environmental conditions, user behaviour (there are humans in the loop) and the function of other building subsystems like the ventilation system and complementary heat sources. For example, the thermal power can increase by 100% during working hours in an office building compared to the heat consumption at night and weekends. The lack of a holistic monitoring system and resulting limited knowledge about substation variables and parameters, and low temporal resolution further complicates the problem to identify faults in these systems with conventional methods. The internal workings of a substation are unknown and the system is a “black box” from a modelling perspective (Isermann, 2006), which renders a district heating system as a system of thousands of black boxes. Therefore, we propose that a combination of domain-specific knowledge about the physics of district heating systems and probabilistic modelling and data analysis is to be used for automated fault detection. This is the approach taken in this work. See Yliniemi (2005) and Pakanen et al. (1996) for additional discussions about the difficulties involved when developing fault-detection methods for regular district-heating substations.

1.2 Hourly district energy data

The rate at which the average temperature of a building can change is limited by the high heat capacity of a building. However, other phenomena can affect the thermal power used by a district-heating substation at much shorter timescales. For example, the use of heated tap water can change rapidly, and control algorithms in automated ventilation systems and complementary heating systems can change set points instantly. This implies that there can be significant *intraday cycles* in the thermal power used by district-heating substations. *Intraweek cycles* are also common, for example in commercial and public buildings where the average daily power can be significantly lower during weekends compared to working days.

If the energy data management system is limited to daily average values, only intraweek cycles can be identified, while the intraday cycles in the power are averaged out. The difference between the hourly minimum and maximum average power of an intraday cycle can be of order 100%, for example in the form of a doubled heating power in public or commercial buildings during working hours compared to evenings and nights. Therefore, the monitoring and analysis of intraday cycles is valuable for the development of information services, for prediction and control of peak loads, for system optimization and for detection of faults. These are some of the reasons why some district energy utilities have upgraded their energy data management systems from monthly or daily values to hourly values. Another motivation is that energy market regulations are expected to become more demanding in the future and utilities prepare for that event when anyway upgrading their data management systems. A still higher sampling rate is motivated if dynamic models are to be implemented, and for estimating the power used for space heating and heating of tap water with a single energy meter (Yliniemi, 2005), but that is presently only feasible in a local or decentralized system due to bottlenecks in the communication systems that are commonly used today.

The methods that are presented in this report are designed for hourly district energy metering data but can be applied also to daily average values, except for the cluster-analysis of intraday load-cycles that is presented in Section 3.1. Cluster analysis methods for daily average values are described in the literature, see Seem (2005), Seem (2007) and Li et al. (2010). Therefore, we focus on hourly values and variables that are typically available in data management systems, see Table 1.1.

Table 1.1. Variables associated with a district energy substation in a data management system.

Quantity	Symbol
Energy or Power (hourly average)	E, P
Flow (hourly average)	m
Primary supply temperature (hourly instantaneous sample)	T_{ps}
Primary return temperature (hourly instantaneous sample)	T_{pr}
Outdoor temperature (hourly estimate, measured elsewhere)	T_{out}

The power is calculated by the data management system from the energy reported by the energy meter. The instantaneous flow is available in some data management systems, but it is of little use because the flow meter typically uses a pulse-based code when communicating with the energy meter and jitter effects can be significant. Therefore, the hourly average flow is preferred because it is more representative for the hourly average power. The primary supply and return temperatures are more problematic because only the instantaneous values are available, which means that a peak load at the time when the hourly temperature samples are fetched by the data management system can render the hourly return temperature non-representative. A similar problem results if the supply temperature varies significantly.

1.2.1 Real-world dataset considered

To demonstrate the methods and heuristics proposed in this report we use a real-world dataset extracted from a district energy management system. The dataset contains about one year of hourly data from 996 district heating substations located in different buildings and constructions in Stockholm, see Table 1.2 for a summary.

Table 1.2. Categories of district heating substations included in the dataset used in this report.

Substation type	Number
Apartment buildings	628
Detached houses	243
Buildings for offices and shops	36
Buildings for apartments combined with other premises	19
Public service	12
Industry and trading	9
Buildings for offices including other premises	6
Special constructions (streets, parking, subway, …)	6
Other buildings (hotels, sports, churches, …)	37
Sum	**996**

The mixture of hourly instantaneous and average values in the data, see Table 1.1, implies that there should be some discrepancy between the true power calculated by an energy meter and the power that we can estimate from the hourly values. How big is that discrepancy? That question can be answered by considering the discrepancy / error introduced by the hourly values

$$P = \frac{1}{\Delta t} \int_{t}^{t+\Delta t} c\, m\ t' \quad T_{ps}\ t' \ - T_{pr}\ t' \quad dt' \tag{2}$$

$$= c\, \overline{m}\ \ \overline{T}_{ps} - \overline{T}_{pr}\ \ + error. \tag{3}$$

Here the flow and primary temperatures with a bar in Equation (3) denote hourly values obtained from the data management system, while Equation (2) describes the true average power that is calculated from the energy received from the meter. The resulting error is illustrated in Figure 1.2 for the whole dataset including one year of data from the 996 substations described in Table 1.2.

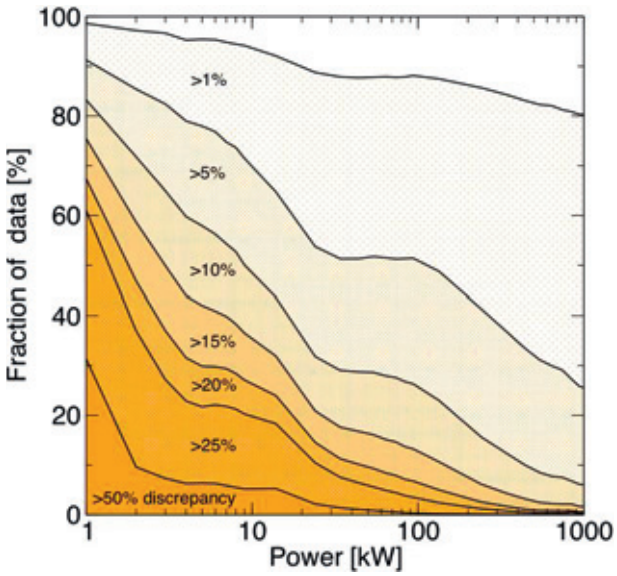


Figure 1.2. Discrepancy between the true power and the power calculated from the hourly primary temperatures and average flow (including faulty data).

The level curves in Figure 1.2 indicate the discrepancy between the power calculated directly from the energy received from the meter (which mostly is correct) and the energy calculated from the hourly primary temperature samples and hourly average flow. For example, at 10 kW about 20% of the data has a discrepancy of 25% or more, and about 50% of the data has a discrepancy of 10% or more. The dataset includes data from a number of documented and suspected faulty substations, which are identified and discussed in the subsequent chapters of this report. Some of these

faults result in outliers that deviate by several orders of magnitude from the normal range, which contributes to the high discrepancies illustrated in Figure 1.2.

The intraday cycles and variation in the power in combination with the errors that are naturally present in hourly district energy data (Figure 1.2) limits the accuracy of methods for fault detection. No matter how accurately we model the expected power there will be deviations from the model, which results from the inaccuracy of the hourly data rather than faults in the system. Therefore, in this work we consider probabilistic models and constraints on the expected deviation of the hourly (or daily) data from the models. Furthermore, ranking of deviations is a key strategy employed because the relatively few large deviations that can be associated with significant faults deserve more attention than the bulk of small discrepancies that are naturally present in the data. Ranking heuristics enables identification of outstanding anomalies with limited labour so that high risks (economic, trust, resources) can be avoided.

1.3 Common faults and symptoms

There are a number of components in a district energy substation that can malfunction or be incorrectly designed or installed, resulting in faulty substation behaviour and incorrect energy meter data. Common components include (Frederiksen & Werner, 2001): cables, valves, flow meter(s), temperature sensors, pressure sensors, pipes, heat exchangers, electronic control system and the electronic energy meter. Incorrect energy meter data can result if any of these components malfunction, which can occur for a number of different reasons. Common faults and issues include (Pakanen et al., 1996; Yliniemi, 2005; Reference group):

- Malfunctioning valves, flow meters and temperature sensors, including faulty voltage references and amplifiers.
- Incorrect installation of substation and associated instrumentation. Such as: incorrect cabling causing electromagnetic compatibility (EMC) issues; splices on flow-sensor cables causing pulse-bounces; incorrect grounding or galvanic isolation; incorrect dimensioning of components, such as valves and flow meters; use of sensors and energy meters that are incompatible.
- Incorrect configuration of meters, sensors and control system.
- Faults or reset of energy meters, for example during a blackout or lightning strike, or drained batteries. Faults in electrical components of energy meters.
- Faults in the communication with the energy meter or time stamping of data.
- Faults introduced during maintenance, or intentionally by customers (fraud).
- Faults introduced during manual recording of energy meter data.
- Fouling or leakage in heat exchangers and pipes.
- Energy meters are misidentified in the management system.

Some of these faults are difficult to detect, for example internal leakage in a heat exchanger, which can result in contamination of domestic hot water with minimal consequences for the hourly energy meter data. This problem can be addressed with

more advanced instrumentation (Isermann, 2011), but it is not feasible to detect using standard energy meter data. Because of the discrepancies inherent in hourly data (see the former section) we focus on the detection of faults that cause significant effects in the data that is commonly available in energy data management systems. In particular, we consider methods for detection of symptoms like:

- Abnormal values of quantities.
- Drift of values over time.
- Excess noise or fluctuations of quantities.
- Constant values of quantities.

Abnormal values and long-term drift can result also when buildings are upgraded or human behaviour changes. Such events have to be distinguished from instrumentation faults manually, but can also be considered as a basis for development of services.

1.4 Related work

Fault detection and diagnosis is an active field of research in many application areas that has stimulated the development of a broad range of methods and heuristics; see Isermann (2005; 2006; 2011) for a general introduction and review, and Katipamula & Brambley (2005) for a review focusing on buildings.

This project and the results that are presented in this report are partially based on the experience from several former projects at the Luleå University of Technology that focuses on fault detection and diagnosis in district heating substations and related instrumentation, see for example Svensson (1996), Carlander (2001), Delsing & Svensson (2001), Jomni (2004), Berrebi (2004) and Yliniemi (2005). The work by Pakanen et al. (1996) and Bergquist et al. (2004) also concerns fault detection in district heating systems and are referenced in this report. The initial developments of the methods that are described in this report are summarized in a conference paper (Sandin et al, 2012). Seem (2005) and Li et al. (2010) develop cluster-analysis methods for the identification of intraweek cycles and weekdays with similar power consumption. That work complements the discussion in this report about identification of intraday and intraweek cycles with cluster analysis of hourly energy meter data. Jota, Silva & Jota (2011) develops an approach for synthesis of daily load shapes that is also based on cluster analysis. Seem (2007) introduces a method for detection and ranking of outliers in the daily average power, which we adopt here for outlier detection and ranking. A novel aspect of the approach taken here is that we model the expectation values of hourly quantities with piecewise linear regression models that are specifically designed for district energy applications, and we use the outlier detection method introduced by Seem (2007) for residual analysis. Piecewise linear regression models have been independently developed and used for modeling of district heating substations by Master's thesis students at the Chalmers University of Technology (Munoz, 2006; Lindquist, 2010), which we learned at the end of this project. A similar piecewise linear regression approach has been developed for

windmills (Forsman, 2011). The piecewise linear regression approach used in these three theses is similar to the one considered here, but we consider hourly data and complement the regression models with probabilistic models for outlier detection and ranking, detection of drift and detection of abnormal quantization.

Other related work include various methods for energy data visualization, see Seem (2007) for references; An approach to detect unexpected changes in the energy efficiency of buildings using algorithmic exploration of district heating billing data (Kiluk, 2012); Analytical models of hourly energy data and fault detection with residual and correlation analysis (Johansson, 2005); Thermal response models of buildings (Armstrong, 2006) and building components (Jiménez & Madsen, 2008), which are interesting for further developments of system optimization methods and dynamical models of district heating substations for fault detection; Methods for short-term forecasting of energy demand (Taylor & McSharry, 2007).

1.5 Aims and scope

This project is motivated by the difficulties experienced among district energy utilities to identify faults in large-scale district energy systems, and to process the numerous false alarms that can result when simplistic methods for fault detection are used. The trend towards use of hourly energy metering data and the growing interest to develop new services and functionality based on high-resolution energy metering data, which enables monitoring and analysis of intraday load cycles, further motivates the development and adoption of new methods for fault detection. This project deals with probabilistic methods for automated anomaly detection and ranking that can be useful for the identification of common faults in existing large-scale district energy systems using information that is available in modern data management systems. In particular, we focus on hourly energy metering data. We do not consider dynamical models and fault detection methods that depend on information that is not commonly available in the energy data management systems, or which can be difficult to apply automatically in a large-scale system. We have tried to identify simple methods and heuristics that can be practically useful in a real-world setup. The proposed methods are studied using hourly data from a population of 996 district heating substations. In our study we have used a Matlab implementation of the methods that are presented in this report, which enables automated analysis of data from several thousand district energy substations. Sample code of key functions needed to implement the methods is provided in the appendix. This does not imply that the methods that are presented in this report are products that are ready to use “as is”, but should rather be considered as proof of concept. It remains to learn which methods that will prove useful and cost-effective in a full-scale implementation. In principle the methods that are discussed in this report may be applicable also for district cooling data, but we do not study that possibility here because the lower primary temperature difference in district cooling applications makes the fault detection problem a more delicate one. Therefore, we propose that these methods should first be implemented and tested in district heating

systems. This report is technical in nature and focuses on methods for anomaly/fault detection. We assume that readers of this report have technical knowledge of district energy substations, energy meters and energy meter data management systems.

2 BASIC METHODS FOR FAULT DETECTION

In this chapter we describe a set of basic methods for the detection of abnormal energy metering data, which are useful indicators of faults that are straightforward to implement. Some of the methods that are described here are already used by the industry (limit checking) and one method (outlier detection) has been developed in this project and is based on the outlier detection approach developed by Seem (2007). This chapter also serves as an introduction to the forthcoming discussion and the methods that are described in subsequent chapters of this report.

2.1 Limit checking

Limit checking with constant thresholds is a basic method that is commonly used for fault detection and diagnosis (FDD); see Isermann (2006) for an introduction. The basic idea is to test whether a measured or derived quantity is within the bounds that are acceptable from a physical, design or safety perspective, or the bounds set by the historical variation of the quantity. An alarm is typically generated when the test fails. Limit checking is straightforward to implement in energy data management systems and it is useful for the detection of some common faults in district energy substations. In particular, the following limit checking tests are recommended and used by some companies (Reference group)

- $T_{ps} \leq T_{max}$, the primary supply temperature should not exceed the maximum supply temperature to the network. This test can fail if there is a fault in the supply temperature sensor, or the related connectors, cabling and electronics. In district cooling this inequality is replaced with $T_{ps} \geq T_{min}$.
- $T_{ps} \geq T_{pr}$, the primary return temperature should not exceed the primary supply temperature. This test can fail if there is a fault in any of the two temperature sensors, or the related connectors, cabling and electronics. In district cooling this inequality is reversed.
- $E_{i+1} \geq E_i$, the hourly energy calculated by an energy meter should not decrease. This test can fail when there is a fault in the power supply of the energy meter that results in a reset of the meter.
- $P \leq \alpha P_{contract}$, the power should not exceed the contracted power by more than some factor α. This test can fail when an instrumentation fault results in abnormal power values. (The difference between contracted power and the actual power is sometimes used also in the calculation of energy cost.)

When data from multiple substations violate some of the limit tests it is possible to rank the substations according to the magnitude of the maximum deviation from the limit. Ranking is commonly used in this work to identify outstanding anomalies. For example, if alarms are generated because some substations in a network violate the

$T_{ps} \leq T_{max}$ test the alarms can be ranked so that substations with high $T_{ps} - T_{max}$ are given priority. The rationale of this approach is that abnormal values that are further away from the expected value or limit are more unlikely to be correct compared to minor deviations (provided that the probability distribution function of the variable or residual is unimodal and not too skew). Given that the resources to monitor and investigate potential faults are limited in practise, the manual efforts should primarily be focused on the most outstanding deviations. Ideally, all potential faults should of course be detected and investigated. The point here is that anomalies should be addressed in the order of potential significance, and a minimum of time should be spent investigating false or insignificant alarms.

2.1.1 Limit checking with linear thresholds

Another test of this type is limit checking with piecewise linear thresholds; see Figure 2.1 for an example. This method is implemented in energy data management software that is commercially available (cf. Enoro AB) and some district energy companies use it for fault detection.

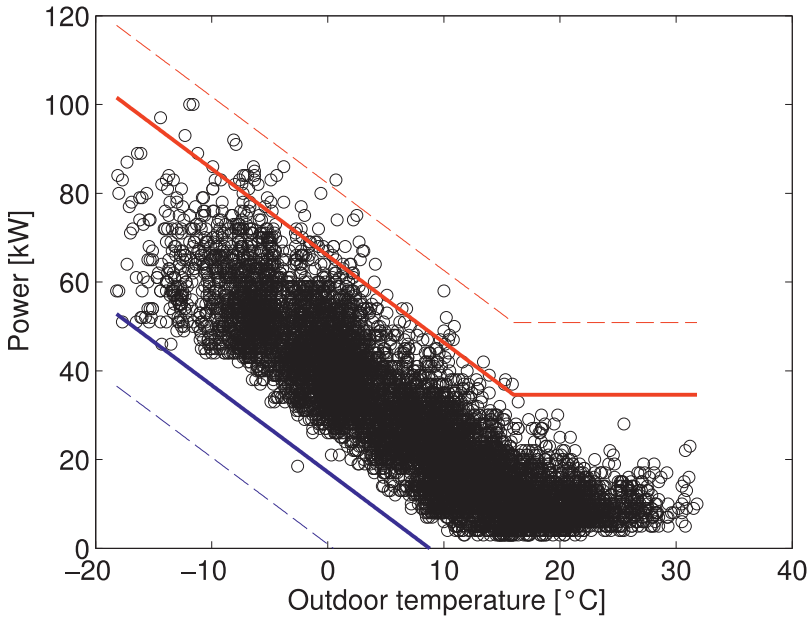


Figure 2.1. Limit checking with linear thresholds of the power. This figure is based on one year of data from an apartment building. The linear thresholds that are illustrated in this figure are located at three standard deviations (solid lines) and five standard deviations (dashed lines) above and below the mean power, respectively.

An alarm is triggered when a value is detected outside the region in-between the lower and upper limits, which typically are set at an empirically determined "reasonable" distance from the temperature-dependent mean power. In this example

the mean power is determined with a least-squares fit of a piecewise linear function, but it can be estimated also with interval-specific averages or median values.

There is a trade-off between the rate of false alarms and the magnitude of variations and potential faults that fall within the acceptable limits. In practice, false alarms have to be accepted when this method is used and the alarms should be ranked in the order of descending maximum deviations to enable identification of outstanding anomalies.

Some substations have significant time-dependent cycles in the thermal load, which for example can result from cycles in the operation of the building ventilation system or cycles in the use of heated tap water. Such cycles can be associated with *branches* in the power profile of substations; see Figure 2.2 for an example.

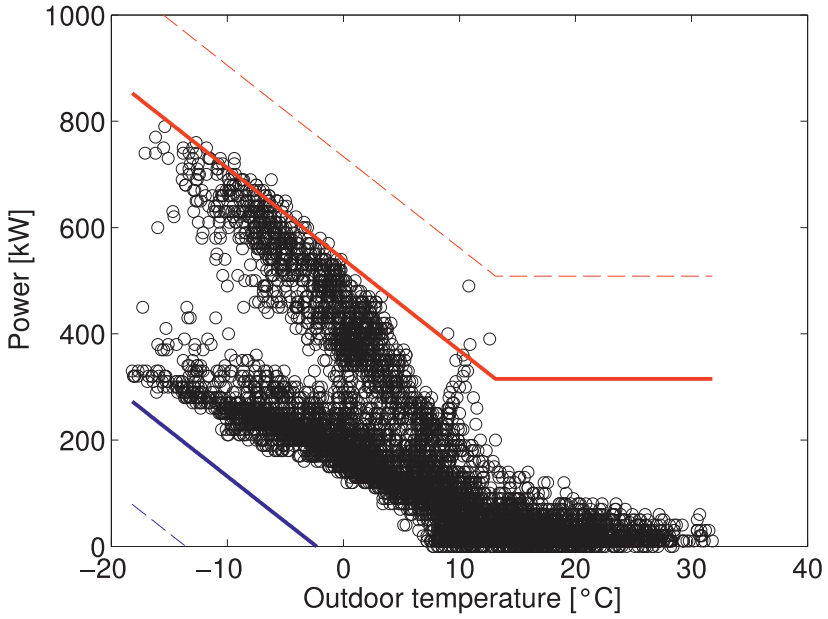


Figure 2.2. Limit checking with linear thresholds of the power for an industrial building with two major branches in the power profile. The linear thresholds are located at three standard deviations (solid lines) and five standard deviations (dashed lines) above and below the mean power, respectively.

In the case of this industrial building the variation of the thermal power during normal operation is substantial and the linear thresholds need to be positioned far away from the actual expected power in order to avoid a deluge of false alarms. This situation can be improved by considering a daily average of the power, which integrates daily cycles, and by using separate linear thresholds for working days and weekends / holidays. More generally, automated methods for the identification of weekdays with similar energy profiles can be used (Seem, 2005; Li et al., 2010). In the next chapter we illustrate how that approach can be extended to hourly data so that substations with branches like those in Figure 2.2 can be identified and modelled.

2.1.2 Limit checking of standard deviation

In addition to limit checking of the mean, μ, of some quantity (like the power or a primary temperature) it is possible to define limits for the standard deviation, σ, of the quantity. The standard deviation holds information about the variation of the quantity. Therefore, it can change when the noise level changes or the value of the variable is constant, which sometimes result from EMC problems, cabling problems and faulty sensors. If the mean and standard deviation before the introduction of the change are denoted by μ_0 and σ_0 and the corresponding values after the change are μ_1 and σ_1 the following cases can be identified

- The mean changes; $\mu_1 = \mu_0 + \Delta\mu$, $\sigma_1 = \sigma_0$.
- The standard deviation changes; $\sigma_1 = \sigma_0 + \Delta\sigma$, $\mu_1 = \mu_0$.
- Both the mean and the standard deviation change.

There are standard tools for the detection of changes of this type, see Sections 7.3-7.5 in Isermann (2006). A basic example is detection of changes in the mean with binary thresholds that are set relative to the standard deviation, which is analogous to the method based on linear thresholds that is described above. This approach works well as long as the change of the mean that is to be detected is large compared to the standard deviation.

Other methods have to be used if the change is less than or similar to the standard deviation. For example, a method for detection of small changes in the mean resulting from signal drift is presented in Chapter 5. Statistical hypothesis tests are common and can be used to detect small changes if the probability distribution of the variable is known. For example, a Student's t-test can be used for detection of changes in the mean and an F-test can be used for detection of a change in the variance (Isermann, 2006). Yliniemi (2005) presents a hypothesis test for detection of changes in the variance (squared standard deviation) of high-pass filtered temperature signals in a district heating substation using a numerically lightweight algorithm than can operate in a microcontroller at 100 Hz sampling frequency. Our experiments indicate that hypothesis tests of this type are less useful with hourly data because the probability distributions of the residuals are complex and varying, and the detailed long-term variation of the mean and standard deviation are difficult to model accurately. Therefore, the methods that we propose for hourly data in the subsequent chapters are based on other types of statistical tests and heuristics.

2.2 Basic method for outlier detection

Methods for statistical *outlier detection* and *scoring* have recently been proposed for the detection of abnormal energy metering data and potential instrumentation faults (Seem, 2007; Li et al., 2010). This approach is a useful complement to the basic limit checking tests that are outlined above. Outlier detection is an appealing alternative to limit checking with linear thresholds because it does not involve the definition of ad-hoc thresholds, and it enables rapid detection of abnormal values. As far as we know,

outlier detection has not yet been implemented and tested in commercial district energy management systems. Therefore, this approach is discussed in some detail in this report and the core functions needed to implement outlier detection are included in the Appendix.

In this section we introduce a basic approach to outlier detection (a more accurate method is described in Chapter 4). The method presented here is a simplified form of the method proposed in Seem (2007), which omits the model of intraweek cycles ("day types") and accounts for seasonal variations with a weekly moving average (Seem, 2005). This approach results in a method that is straightforward to implement, but which is less sensitive to outliers when there are significant intraweek and intraday cycles in the thermal load. This point is further discussed and a solution is presented in the following chapters.

The thermal power varies significantly with the outdoor temperature; see Figure 2.1 for an example of a typical trend. Therefore, in order to enable identification of abnormal power values it is necessary to model the temperature-dependent variation. In principle, there may also be effects of seasons, sun, wind and time-varying human behaviour on the thermal load. A weekly moving average of the thermal power is a basic estimate for the temperature-dependent, seasonal variation of the power (Seem, 2005); see Figure 2.3 for an example. It is necessary to average over one week because intraweek and intraday cycles in the thermal load can otherwise result in artificial short-term cycles in the moving average. A weekly moving average, MA_i, is defined as the mean of the last $n = 7 \times 24 = 168$ hourly power values,

$$MA_i = \frac{P_{i-(n-1)}+\ldots+P_{i-2}+P_{i-1}+P_i}{n}, \tag{4}$$

or the corresponding average of an equal number of values that are selected from both sides of a central value

$$MA_i = \frac{P_{i-\,n/2}\,+\cdots+P_{i-2}+P_{i-1}+P_i+P_{i+1}+P_{i+2}+\cdots+P_{i+\,n/2}}{n}. \tag{5}$$

The latter, centralized moving average is preferred when dealing with historical data because the former definition results in a time-offset of the estimated value. In both cases, the moving average can be calculated iteratively by subtracting the oldest term and adding the most recent term, for example

$$MA_i = MA_{i-1} - \frac{P_{i-n}}{n} + \frac{P_i}{n}. \tag{6}$$

The result of this basic seasonal de-trending approach is illustrated in Figure 2.3.

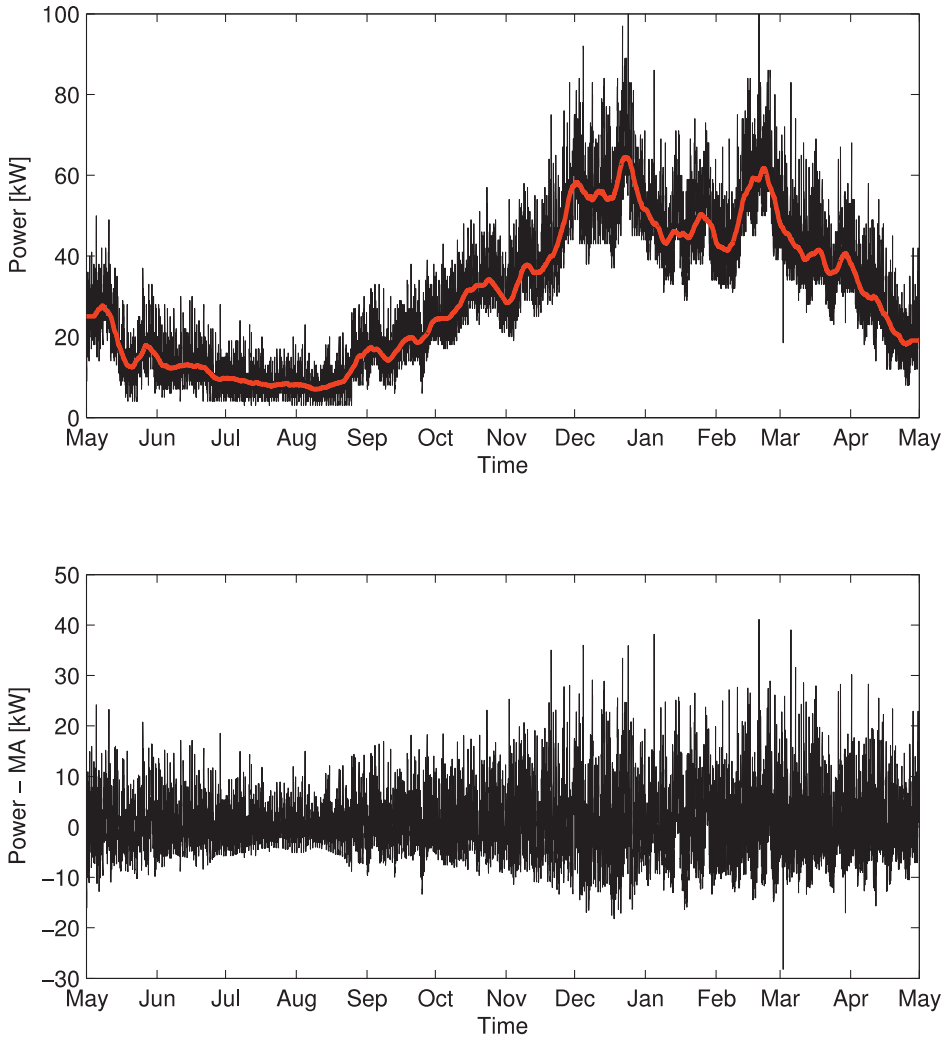


Figure 2.3. Seasonal detrending of the power with a weekly moving average (MA). The temperature-dependent variation of the power is approximated with a weekly moving average, which is invariant to intraweek and intraday cycles in the power. The upper panel shows the hourly power values over one year (black line) and the weekly moving average (red color). The lower panel shows the de-trended power, after subtraction of the weekly moving average. This is the same substation as that illustrated in Figure 2.1.

After subtraction of the moving average, a standard outlier detection test can be applied to detect abnormal power values. The generalized extreme studentized deviate (GESD) test (Rosner, 1983) is recommended when the number of potential outliers is unknown because it works well under a variety of conditions (Iglewicz & Hoaglin, 1993) and it has been proposed for the detection of outliers in energy meter data (Seem, 2007; Li et al., 2010). The mathematical details of this outlier test are

described in Chapter 4 and an implementation of the basic outlier detection method is provided in the Appendix; see the functions named `basic_test` and `gesd`. In principle, the idea is that the de-trended power values are expected to stay within a few standard deviations from zero. The GESD test provides a formal answer to the question whether a certain deviation from the average is likely given that the variable is approximately normally distributed with some given standard deviation, or whether the deviation should be considered as an outlier. Figure 2.4 illustrates some outliers in district heating data that were detected with this method.

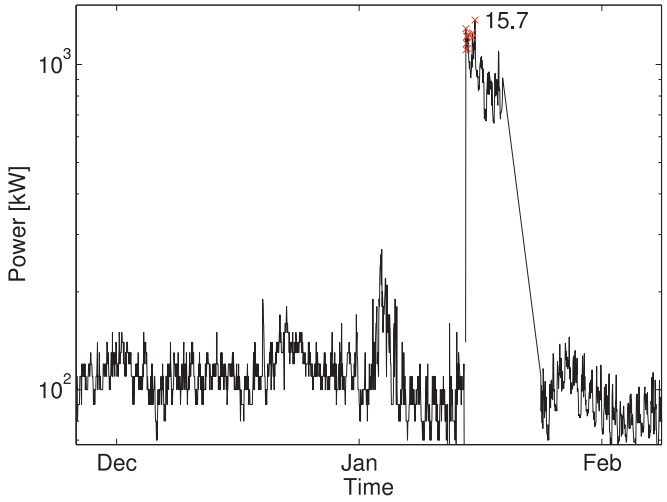


Figure 2.4. Outliers in the thermal power of an apartment building that were detected with the basic method. The ten most significant outliers are tagged (crosses) and the maximum deviation is 15.7 standard deviations. A deviation of fifteen standard deviations is extremely unlikely under normal operation, and the probability of having several such outliers in a sequence is practically zero. The outliers are caused by a fault in the instrumentation. The energy meter, including primary temperature and flow sensors were replaced on the 19th of January.

This figure was produced with the function named `basic_test` in the Appendix. By accounting for intraweek (Seem, 2007; Li et al., 2010) and intraday (Sandin et al., 2011) patterns in the power using the methods described in the following chapters the variance of de-trended data (residuals) can be further reduced, which improves the sensitivity and reliability of the outlier detection test.

2.2.1 Ranking of outliers

Outliers can be ranked depending on how much they deviate from the expected value. One way to do that is to score the outliers with the corresponding number of standard deviations, σ, a so-called standard score, or *Z score*. For example, a Z score of 10 implies that the outlier deviates from the expected or average value by ten standard

deviations. The details and mathematical definition of Z scores are further discussed in Chapter 4. Qualitatively, for a normally distributed random variable about 68% of the values are within one standard deviation from the average (Z scores less than 1); about 95% are within two standard deviations (Z scores less than 2); and about 99.7% are within three standard deviations (Z scores less than 3). Therefore, most of the observed values of a variable that is approximately normally distributed should be within three standard deviations from the average. This is a rule of thumb known as the "3-sigma" rule. In practise, the empirical results of the basic outlier detection method that are presented in the next subsection shows that Z scores up to 10 are common on an annual basis, and scores exceeding 10 are abnormal.

2.2.2 Test results

We apply the basic outlier detection method to one year of hourly power values for the 996 substations in the test set, see Chapter 1. For each substation we calculate the maximum magnitude of the Z scores. The result is presented in Figure 2.5, which illustrates how many substations (in %) that have an outlier with a maximum magnitude of the Z score of some value. Outliers are detected in the power data of most substations with this method, but only a subset of the substations has exceptionally high Z scores. The faulty substation with a maximum Z score of 15.6 that is illustrated in Figure 2.4 is found at position 28 in the top-list, which means that there are 27 substations in this data set that have higher maximum Z scores. There are 50 substations that have no outliers according to the basic method. There are 70 substations that have a Z score above 10, 30 substations with a Z score above 15, and 19 substations with a Z score above 20. The six substations with the highest Z scores are illustrated in Figure 2.6, and the six substations at positions 10-15 in the top-list are illustrated in Figure 2.7.

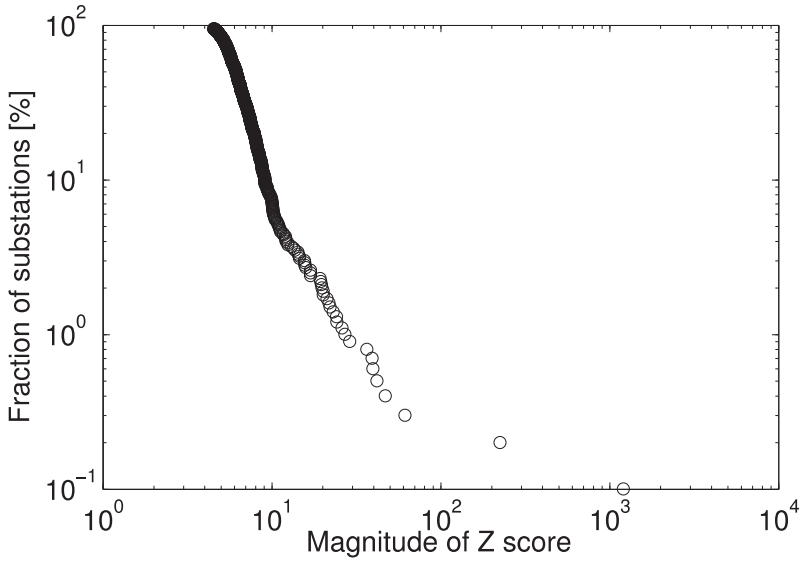


Figure 2.5. Maximum magnitude of Z scores for the 996 substations in the test set. This figure is based on one year of hourly data. The substation in Figure 2.4 with a maximum Z score of 15.6 is represented by the circle at position 28 from the right-hand side, which means that there are 27 substations in the test set that have outliers with higher Z scores.

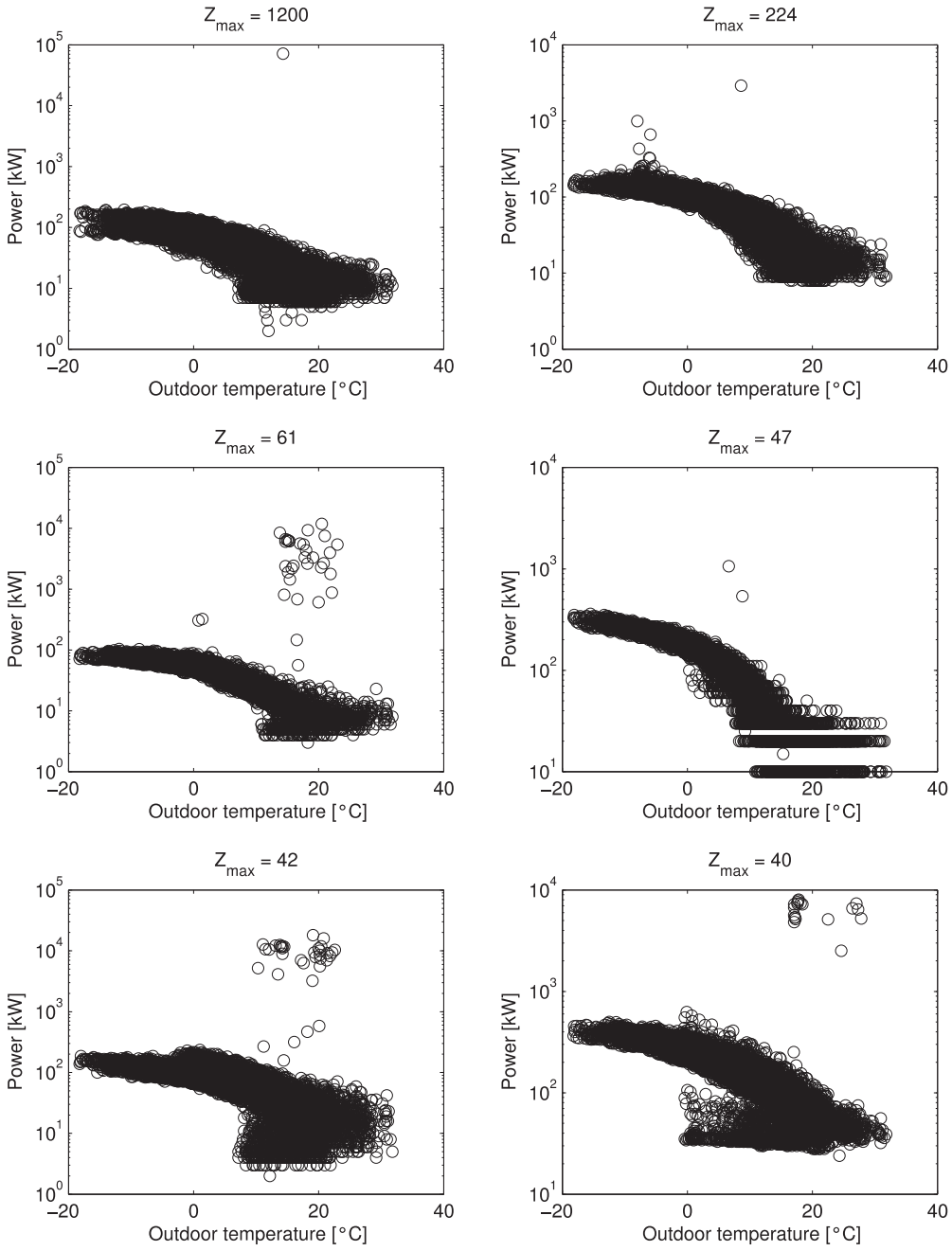


Figure 2.6. The top-six substations in the test set with the highest Z scores of outliers identified with the basic method. Faults are suspected or confirmed in all six cases. For example, the substation with a maximum Z score of 61 has been rebuilt due to problems with the energy meter and communication. The substation with a maximum Z score of 42 had a faulty energy meter that was replaced.

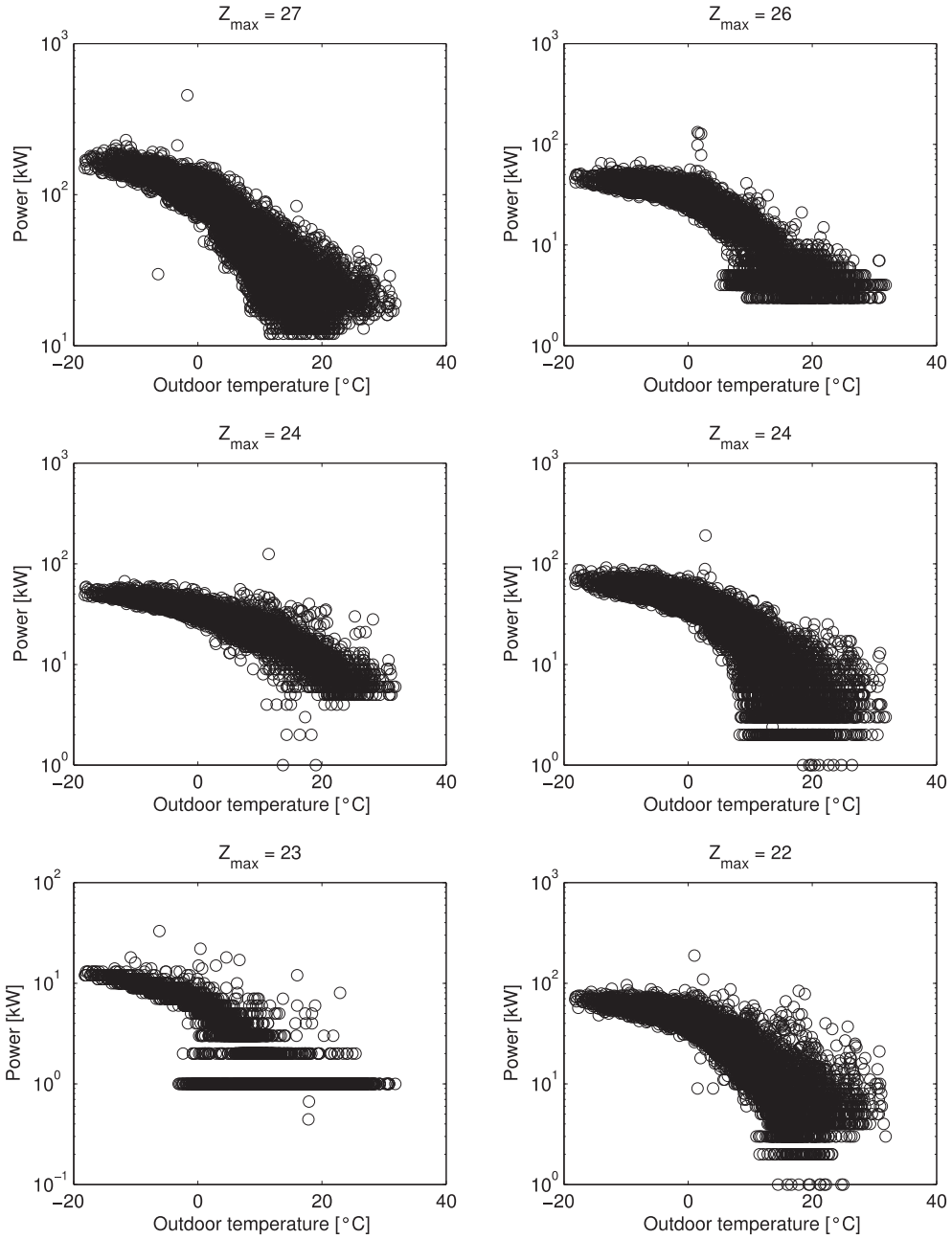


Figure 2.7. Substations in the test set with high Z scores of outliers identified with the basic method. These six substations are found at positions 10-15 in the top-list of substations with high Z scores, see Figure 2.5.

3 PROBABILISTIC MODELS

The moving-average estimate of the thermal power of a district-heating substation that is introduced in the former chapter is useful for basic outlier detection, but it is not an accurate model of the power when there are significant intraday or intraweek cycles in the thermal load. Such cycles naturally appear in some buildings. For example, the ventilation system can be switched off at night and during weekends in some buildings, which results in intraday and intraweek cycles in the thermal power needed for space heating. This is a common phenomenon in industrial buildings, office buildings and public buildings like schools. Cycles in the use of heated tap water are also common. Another limitation of the moving-average estimate is that it is continuously updated to match the actual mean power used by the substation, which makes it useless for detection of long-term changes in the thermal load. For example, long-term changes can result when buildings are upgraded, when ventilation systems or complementary heating systems are modified, and because of faults in the instrumentation that lead to signal drift. Therefore, a better model of the thermal power is needed. Models of other variables that are commonly available in district energy data management systems, in particular the primary temperatures and the flow are also useful for fault detection and diagnosis.

In this chapter we describe how such models can be automatically generated using statistical tools. In particular we use piecewise linear regression, bimodality analysis, cluster analysis and correlation analysis (Murphy, 2012). The resulting models are used as input to the probabilistic outlier detection methods that are described in Chapter 4 and the drift detection method that is described in Chapter 5. We use probabilistic modelling rather than an approach based on dynamic equations of buildings and substations because of the low sampling rate of one sample per hour considered here and the few variables that are known. A probabilistic approach is also motivated by the need to model large populations of buildings and substations, which are affected by human behaviour, weather conditions, and the complex interactions with other systems like various types of ventilation systems. Effects like these can be modelled approximately with data-driven probabilistic methods, but would be more difficult to model automatically with dynamic equations.

3.1 Thermal power

A fault in the instrumentation that results in incorrect thermal power directly affects the energy cost and billing. Therefore, the power is a key quantity for fault detection. It is important that faults resulting in abnormal power values are detected as soon as possible. The thermal power, P, is defined here as the hourly mean power obtained from the time-derivative of hourly samples of the energy, E, calculated by an energy meter, $P = \dot{E}$. The power is more convenient to analyse than the energy because the power is proportional to the thermal load, while the energy is a monotonic function.

The most important variable affecting the thermal load and power used by a district energy substation is the outdoor temperature. In particular, the need of thermal power for space heating varies significantly with time in countries with seasons, where the outdoor temperature varies from sub-zero temperatures in the winter to indoor temperatures or higher in the summer. This effect is evident in the power data that is illustrated in Figures 2.1 and 2.3, and it is a consequence of two main principles; the hourly thermal power needed for space heating in a building is approximately proportional to the difference between the outdoor and indoor temperature; and the substation control system uses the local outdoor temperature as reference to calculate the control set point for space heating. The control system typically uses a linear relationship between the set point and the local outdoor temperature, within some upper and lower bounds. Space heating is typically turned off when the outdoor temperature is comparable to the indoor temperature; for example at 2–3°C below the indoor temperature, which translates to outdoor temperatures of 16–17°C or higher (this is the basis for the traditional concept of "degree days"). There is also an upper limit to the power because the secondary supply temperature is kept below some value for safety reasons, for example 60°C, and the heat exchanger(s) anyway have limited capacity. Therefore, the relationship between the average power and the outdoor temperature is expected to be a *piecewise linear* function; see Figure 3.1 for an example.

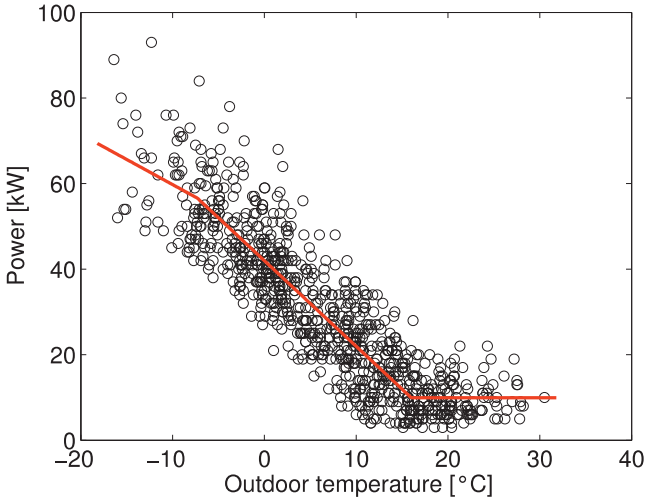


Figure 3.1. A piecewise linear model of the hourly average power versus the outdoor temperature for an apartment building in Stockholm. The model is fitted with adaptive piecewise linear regression using ARESLab (see the Appendix) and four linear segments separated by three breakpoints, out of which two breakpoints are clearly visible. The regression model is fitted to one year of data but only 10% of the data is displayed. There is an evident knee in the trend at about 16°C, which is related to the onset of space heating.

The functional relationship between the outdoor temperature and power can be more complicated than that illustrated in the figure above for several reasons; intraday and intraweek cycles in the thermal load can be significant, this point is further discussed below; some control systems use a piecewise linear relationship between the set point and the outdoor temperature to enable further customization or optimization of space heating and energy use; the supply temperature to a district-heating network is controlled with weather forecasts and varies with the outdoor temperature, which affects the function of the substation. For example, the supply temperature to a district heating substation can be about 70°C in the summer and above 100°C in the winter. For these reasons, it is motivated to use a piecewise linear model with several breakpoints when automatically constructing models for a whole population of substations. A piecewise linear model with one breakpoint at the onset of space heating (for example at T_{out} =16–17°C) is a fairly good approximation for some substations, but in general the model can be improved by adding a few more breakpoints. We return to this point below, after a discussion about modelling of intraday and intraweek cycles.

3.1.1 Schedule of intraday and intraweek cycles

Intraday and intraweek cycles in the thermal load are common. For example, the ventilation system can be re-configured or switched off at night and during weekends in some buildings, which results in intraday and intraweek cycles in the thermal power needed for space heating. This is a common phenomenon in industrial buildings, office buildings and public buildings like schools. Another common cycle is related to the varying use of heated tap water in apartment buildings. Cycles of these types can cause significant variations in the thermal load; see Figure 3.2 for an example. Therefore, the model of the expected hourly power should account for eventual cycles in the thermal load. Note that the outdoor temperature in Figure 3.2 reaches a minimum on Friday evening and that the trends before and after this point are visible also in the power. The average thermal power follows the trend of the outdoor temperature. Note that the intraday and intraweek cycles in the thermal load affects the power of order 100% at the timescale of one hour. Therefore, any model that is based on hourly data (whether probabilistic or dynamic) that fails to account for these cycles will be incorrect. Figure 3.3 illustrates a one-year power profile of the substation that is discussed in the example above.

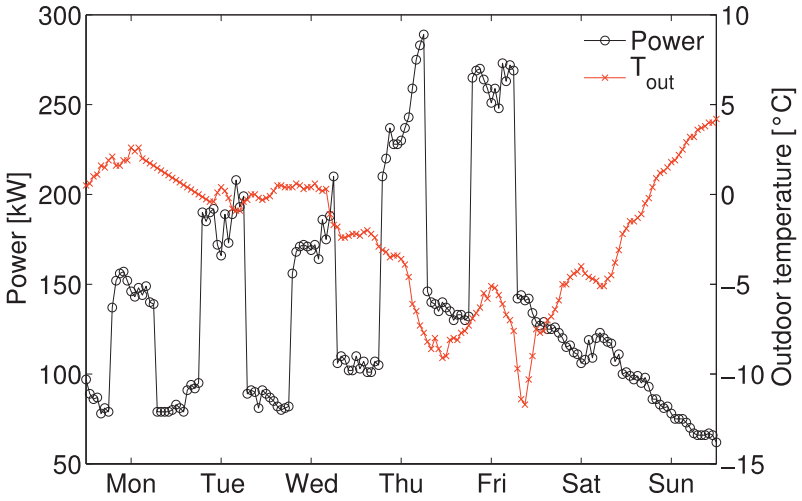


Figure 3.2. Thermal power and outdoor temperature versus time for an office building in Stockholm during one week in January, 2011. There are evident intraday and intraweek cycles in the power. The thermal load is higher during office hours, between 7:00 and 18:00 from Monday until Friday.

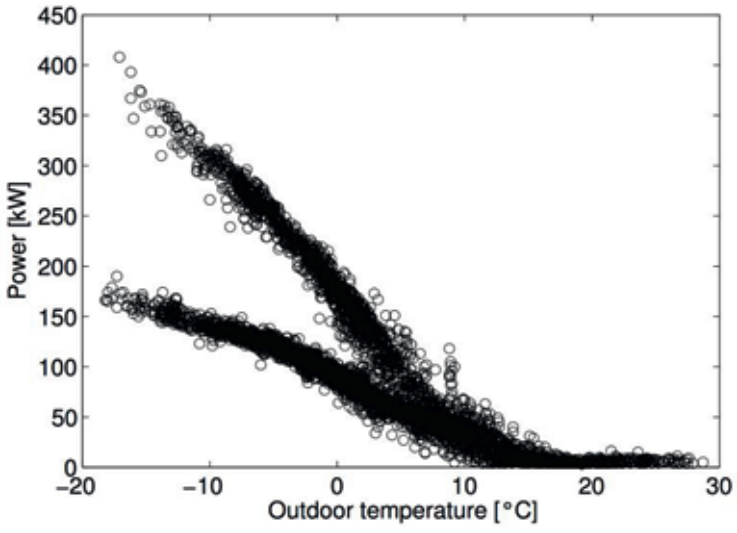


Figure 3.3. One year of power data versus the outdoor temperature for the office building that is illustrated in Figure 3.2. The difference between high and low thermal loads is evident at outdoor temperatures below zero.

Former work on outlier detection in energy meter data (Seem 2005; Seem 2007; Li et al., 2010) focuses on daily averages, which means that intraday cycles are averaged out and that only intraweek cycles remain. In this context methods for automatic identification of weekdays with different thermal load (so-called "day types") have been developed. These methods are based on cluster analysis of average and peak daily energy use (Seem 2005; Seem 2007) and cluster analysis of average and peak daily energy use combined with autoregression coefficients (Li et al., 2010). In this work we consider hourly data, which implies that we need to model the intraday cycles in addition to the intraweek cycles. For initial steps in that direction see Jota et al. (2011), Kiluk (2012) and Sandin et al. (2012). In addition, we are aiming to develop a model that can be automatically used with a population of tens of thousands of substations. Therefore, the model should be simple, comprehensible and easy to modify in the event that the automatically calculated results are unexpected or incorrect. For these reasons we propose to model the cycles in terms of a weekly schedule of the thermal load, which can be automatically calculated using cluster analysis (Sandin et al., 2012); see Figure 3.4 for an example.

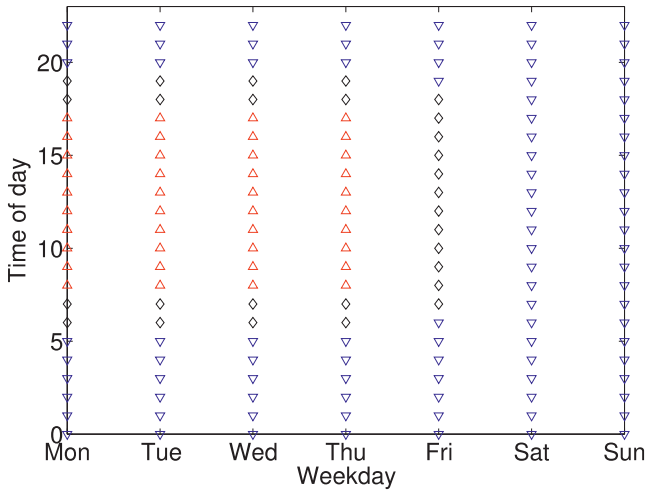


Figure 3.4. Schedule of intraweek and intraday cycles in the thermal power of an office building. This is the same building as that illustrated in Figures 3.2 - 3.3. The high power (triangles, pointing upwards) and low power (triangles, pointing downwards) cycles are aligned with the office hours. The third symbol (black diamonds) should be interpreted as "either high or low".

A schedule of this type is straightforward to understand and modify. It will fail in some cases, for example at public holidays, but custom rules can be defined manually for such exceptional cases. Schedules of this type are also useful as a diagnostic tool that can be used to investigate how cycles change over time. Next we describe how schedules of this type are calculated. The code used to create this and other schedules

displayed in this report is provided in the Appendix, see the functions named `std_power`, `bimodality` and `power_schedule`.

3.1.2 Standardized power

It is tricky to analyse cycles in the time sequence of power values directly because of the outdoor temperature dependence, see Figure 3.2 for an example, and because of the varying amplitude of the cycles between different substations. Using basic methods it is also difficult to fit a piecewise linear function to the power profile *before* we have identified and analysed eventual cycles in the power. Therefore, we *standardize* the power variable using a temperature-specific mean power and standard deviation of power (note that the term "standardized power" introduced here is a statistical term, which have nothing to do with ISO standards)

$$P_s = \frac{P - E_T[P]}{\sigma_T[P]}. \tag{7}$$

The temperature-specific mean power, $E_T\ P$, and standard deviation of power, $\sigma_T[P]$, are calculated by considering power values that are measured within a given interval of outdoor temperatures, which should be narrow so that the temperature-dependence of the power within that interval is negligible. We define the length of that interval as 1°C and divide the power data in intervals of one degree Celsius. For example, power values measured at an outdoor temperature between -5°C and -6°C are combined when calculating the corresponding standardized temperature, P_s, according to the equation above. Similarly, power values in the interval [-7°C, -6°C) are combined when calculating the corresponding P_s values, and so on.

The result of this algorithm is illustrated in Figures 3.5 – 3.6, which displays one year of data from the substation that is displayed in Figures 3.2 – 3.4, with the constraint that $T_{out} \leq -5$°C (about 12% of the annual data fulfil that constraint). An upper limit on the outdoor temperature is used because cycles are less evident at outdoor temperatures near and above 0°C. If data for higher outdoor temperatures are included the cycles in the standardized power are less evident and may prevent identification of the intraday and intraweek cycles. Therefore, the analysis of cycles that is discussed here is not applicable if the effect of the cycles is small compared to the standard deviation of the power. However, if there are no evident cycles in the power it is sufficient to model the outdoor-temperature dependence, so this limitation is not an issue.

Note that there is no evident effect of the outdoor temperature on the standardized power because P_s varies primarily with the weekday and time of day. Also note that the cycles that appear in Figure 3.5 correspond to the cycles that are illustrated by the schedule in Figure 3.4. Next we describe how cycles like these can be automatically identified with bimodality analysis and how a weekly schedule can be generated using cluster analysis.

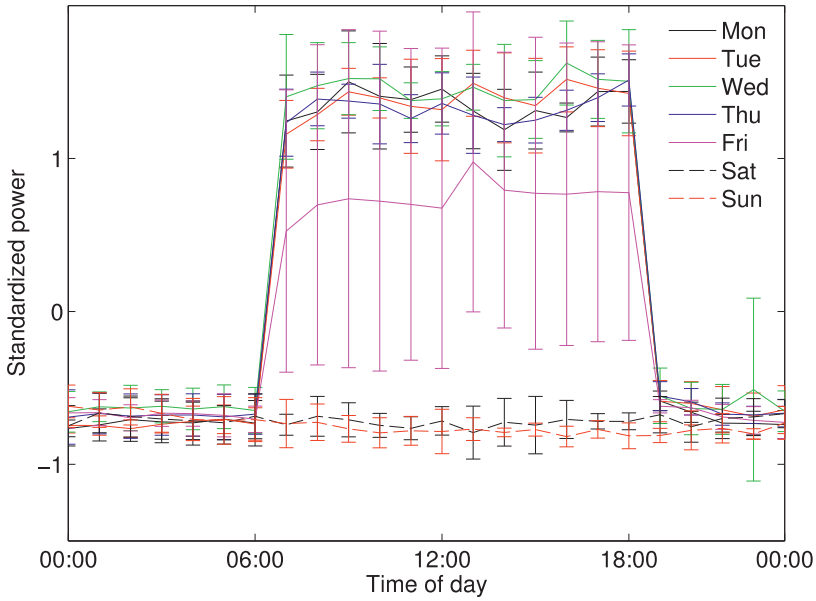


Figure 3.5. Average standardized power, P_s, for the office building that is illustrated in Figures 3.2 – 3.3. This figure is generated from one year of hourly data with the constraint that the outdoor temperature is less than -5°C. Error bars denote standard deviations.

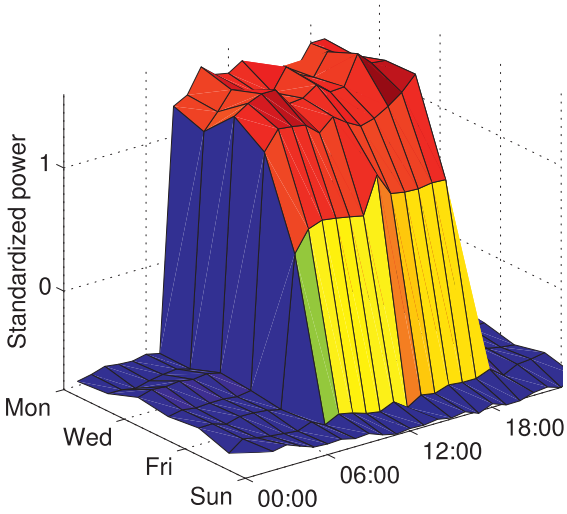


Figure 3.6. Surface plot of the average standardized power for the same office building as that illustrated in Figure 3.5.

3.1.3 Bimodality coefficient

The intraday and intraweek cycles in the thermal power that are illustrated in Figures 3.2 – 3.3 result in a standardized power that changes significantly with weekday and time of day according to Figure 3.5. The standardized power can be represented in the form of a histogram, which is a discretized form of the probability density function of the standardized power variable, P_s; see Figure 3.7. The histogram shows that the distribution function is *bimodal*, which means that the probability density function has *two* distinct peaks (local maxima). In general, a probability density function that has more than one local maximum is said to be *multimodal*. In contrast, a *unimodal* probability density function has only one maximum. Bimodal distributions often arise as a mixture of two different unimodal distributions, which is a natural way to think about the problem considered here.

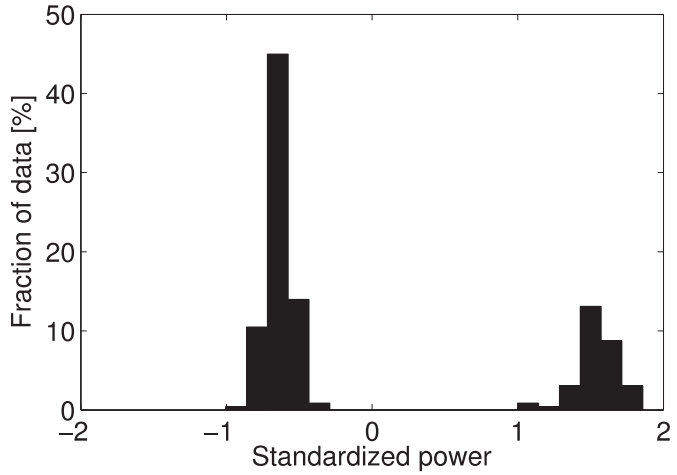


Figure 3.7. Histogram with 20 bins of standardized power for an office building with intraday and intraweek cycles in the thermal power. This result can be compared with Figure 3.5, which is generated with data from the same building.

A bi- or multimodal probability density function of the standardized power indicates that there are systematic cycles in the power. We want to estimate whether assuming that there are cycles in the power is more appropriate than assuming no cycles. Therefore, we introduce a statistic that indicates whether a distribution is multimodal or unimodal. This is a general problem that is studied in several different fields. For example, a bimodal distribution can indicate that there are novel or particularly interesting aspects in the data that are not accounted for by a model (this is for example the case in the context of gene expression analysis). Summary statistics like the mean, median and standard deviation can be misleading for bi- and multimodal distributions, which is another motivation for the use of bimodality analysis.

There are several different techniques for the detection of bimodal and multimodal distributions, but no universal method that is suitable for all problem domains. Some

tests are based on the outcome of a cluster analysis algorithm, while other statistics are functions of the data. For example, the kurtosis of the probability distribution function is a useful statistic that is used in several methods for bimodality analysis. We propose that a combination of the kurtosis and skewness known as the *bimodality coefficient* (BC) is used because it is simple to calculate, it varies between zero and one forming a comprehensible statistic, and our empirical results indicate that it efficiently separates substations with evident cycles in the load (high BC) from substations with no evident cycles (low BC). The bimodality coefficient is defined as

$$BC = \frac{\gamma[P_s]^2+1}{\kappa[P_s]}, \tag{8}$$

where P_s is the standardized power, γ is the skewness and κ is the kurtosis (not the excess kurtosis). The BC of the standardized power illustrated in Figure 3.7 is 0.94, which is a high value, in agreement with the evident cycles in the power. Our empirical results suggest that a BC below ~0.6-0.65 indicates that there are no significant cycles in the power, while higher values motivates further analysis (cluster analysis) of eventual cycles in the standardized power. The BC for the population of 996 substations in the test set is displayed in Figure 3.8.

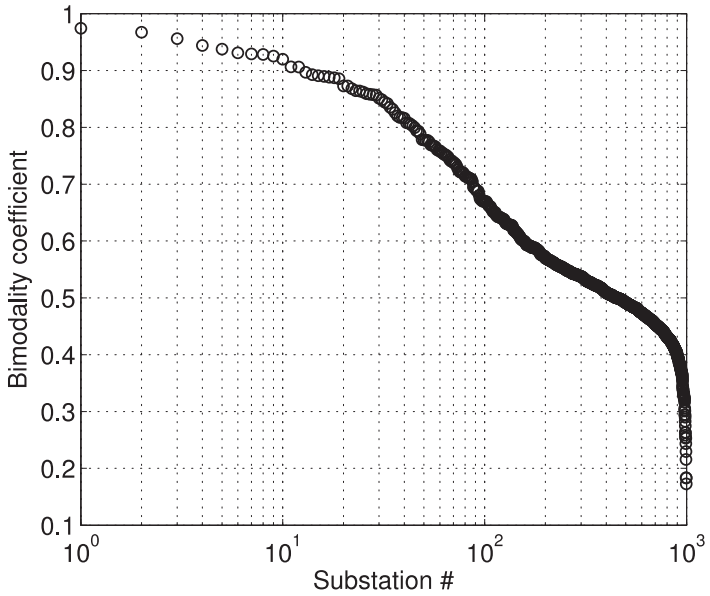


Figure 3.8. Bimodality coefficients for the population of 996 substations in the test set. In practice we find that substations with bimodality coefficients above 0.6-0.65 have significant intraday and/or intraweek cycles in the power.

We find that there are 12 substations in the test set with a BC exceeding 0.9, 44 substations with a BC > 0.8, 87 substations with a BC > 0.7 and 156 substations with a BC > 0.6. This implies that there are evident cycles in the power in about 10-15% of

the substations, which needs to be further analysed with cluster analysis. The remaining 85-90% of substations can be modelled with one regression model. Power profiles of six different substations with varying BC are displayed in Figure 3.9.

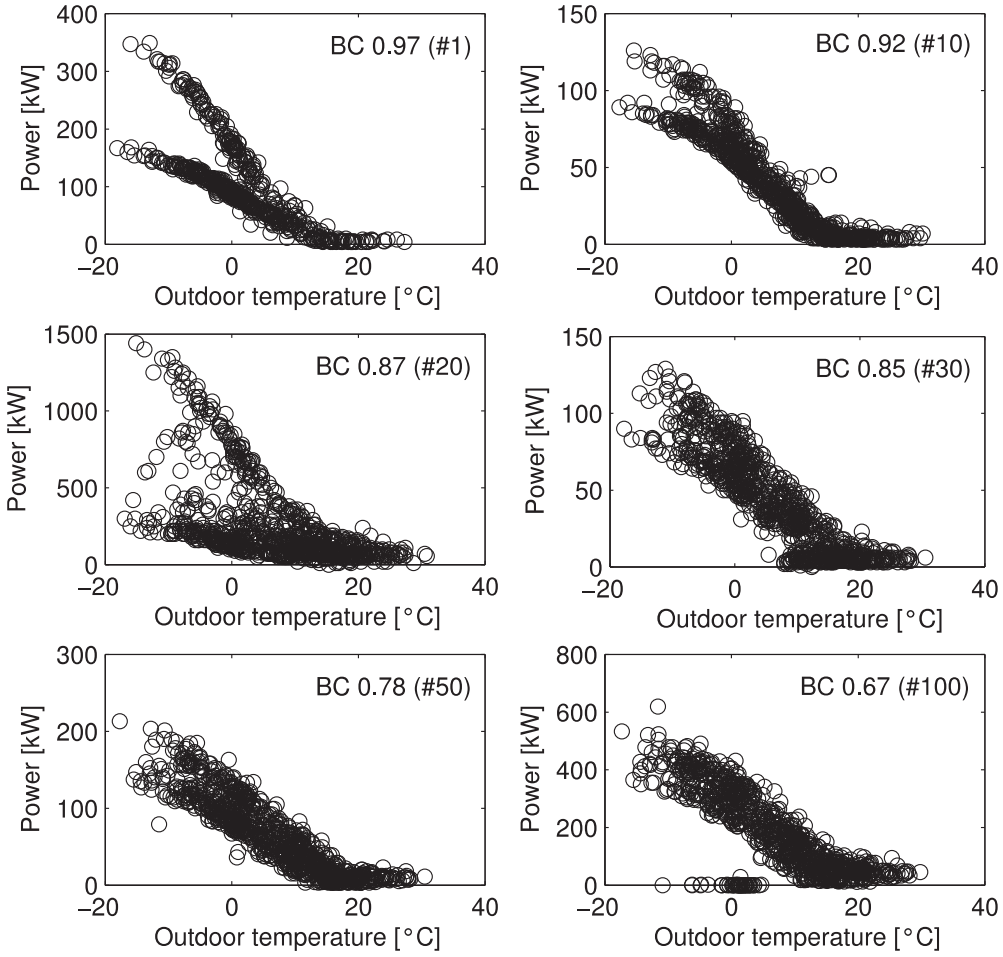


Figure 3.9. Six power profiles of substations with different BC. The qualitative trend is that substations with a high BC have evident cycles in the power, while substations with a low BC have not. The number indicated in the parenthesis of the label in each paned denotes the substation number in Figure 3.8.

We have not found a throughout analysis and discussion of the bimodality coefficient in the peer-reviewed literature. Negative excess kurtosis is commonly used as an indicator of bimodal distributions, but it has some limitations. For example, a skewed bimodal distribution can have high excess kurtosis, which intuitively motivates the definition of the bimodality coefficient in terms of the kurtosis and the skewness. The story is that Warren Sarle suggested the formula and that it originates from ideas taught in a statistics class at the University of Florida, but there is no publication record to confirm that information that we are aware of. The BC is 1 for a Bernoulli

distribution (maximum bimodality), 1/3 for a normal distribution, and near zero for distributions with heavy tails. The BC is related to a formula suggested in 1894 by Karl Pearson, who was the first to propose a procedure for testing whether a distribution can be decomposed into two normal distributions. This anecdotal discussion is included here as a historical reference. The motivation for proposing the bimodality coefficient for this particular application comes from our empirical experiments, which demonstrate that substations with a high BC of the standardized power have evident cycles in the power, while the large majority of substations with low BC do not. Empirically we find that cycles are barely visible at a BC of about 0.6, sometimes visible at a BC of about 0.65 and can exceed 0.9 for substations with major cycles like those illustrated in Figure 3.3 and Figure 3.5. Next we describe how cluster analysis can be used to identify the intraday and intraweek cycles for substations with a high BC, and how to calculate the weekly schedule describing the cycles in the power.

3.1.4 Cluster analysis

Cluster analysis is a natural approach to analyse eventual time-dependent cycles in the power. Former work has mainly focused on the daily average power (Seem 2005; Seem 2007; Li et al., 2010), which means that intraday cycles are averaged out and that only intraweek cycles remain. In that context methods for identification of weekdays with different thermal load have been developed. These methods are based on cluster analysis of average and peak daily energy use (Seem 2005; Seem 2007) and cluster analysis of average and peak daily energy use combined with auto-regression coefficients (Li et al., 2010). Initial developments in the direction of modelling also the intraday cycles can be found in Jota et al. (2011), Kiluk (2012) and Sandin et al. (2012).

Empirically we find that most substations with evident intraday and intraweek cycles can be classified with three clusters, corresponding to *high power*, *low power* and *mixed power*. The latter category represents weekdays and time of day when the power can be either high or low, or something in-between high and low, which sometimes is the case near transitions between high and low power levels. Data belonging to the mixed class is more common for substations with less evident cycles and low BC statistics, while substations with high BC and evident cycles in the power typically have relatively few occurrences of the mixed category. A cluster analysis approach can be used to test whether dividing the data in three categories (high, mixed and low power); two categories (high and low power); or one category is most appropriate. This approach works well for the majority of the 996 substations that are included in the test set used in this study, but there are exceptions to this pattern. For example, one substation that heats a public building with an outdoor swimming pool needs exceptionally high thermal power in the summer when the pool is heated, while the power profile otherwise resembles that of an ordinary building.

A straightforward method to assign categories to the different levels of thermal load represented in the cycles of standardized power is to use the well-known k-means algorithm. In general this algorithm operates in the following way. Let $x_1, \ldots, x_n$ be a set of observations (the average standardized power versus the $7 \times 24 = 168$ hours of one week). Given $k < n$ initial centre points of the potential clusters the k-means algorithm partitions the data into k clusters by minimizing the sum of squared distances to the cluster center points. Formally, the problem is to minimize

$$\sum_{j=1}^{k} \sum_{x \in C_j} (x - x_j)^2, \tag{9}$$

where $C_j, j = 1, \ldots, k$ are clusters with centers

$$x_j = \frac{1}{n_j} \sum_{\in C_j} x. \tag{10}$$

The k-means algorithm operates iteratively by alternating between assigning data points to clusters based on their distances to cluster centres, and updating the centres based on the cluster assignments. For further details, see Section 11.4.2 and Algorithm 11.1 in Murphy (2012), and the implementation in the Appendix of this report named `power_schedule`. The k-means algorithm is sensitive to the initially selected cluster centres. One general approach to define the initial centres is to select k data points at random from the data set, and to select the subsequent centres from the remaining points with a probability that is proportional to the squared distance to the closest cluster centre point. This approach is known as k-means++, see Section 11.4.2.7 in Murphy (2012) for further information.

Here the problem to select initial centre points is simple because the data is univariate and we are explicitly searching for clusters of high- and low standardized power. Therefore, we select the initial cluster centres as the 10^{th} and 90^{th} percentiles of the standardized power. If the result of the k-means algorithm is that all data points belong to one cluster the algorithm exits and it is concluded that there are no cycles in the power, which implies that no weekly schedule is generated and that only one regression model is fitted to the power profile. If two clusters are identified, a third centre point is defined as the mean of the corresponding two cluster centre points and the cluster analysis is repeated with the resulting three initial centre points. The purpose of this second cluster-analysis step is to determine whether there are data points in-between the high- and low power clusters that should be categorized as intermediate values (the "mixed" class introduced above). Optionally, adjacent hours during a week that correspond to a transition from high to low power, or vice versa, can be re-classified as mixed power to avoid misclassification, for example because of jitter in the time stamps of hourly metering data. This third re-classification step is used when calculating the weekly schedule illustrated in Figure 3.4. The cluster analysis procedure provides a category label (high-, mixed-, or low power) for each of the 168 hours of the week. These categories can be displayed in the form of a

weekly schedule with three possible categories for each weekday and time of day, see Figures 3.10 – 3.11.

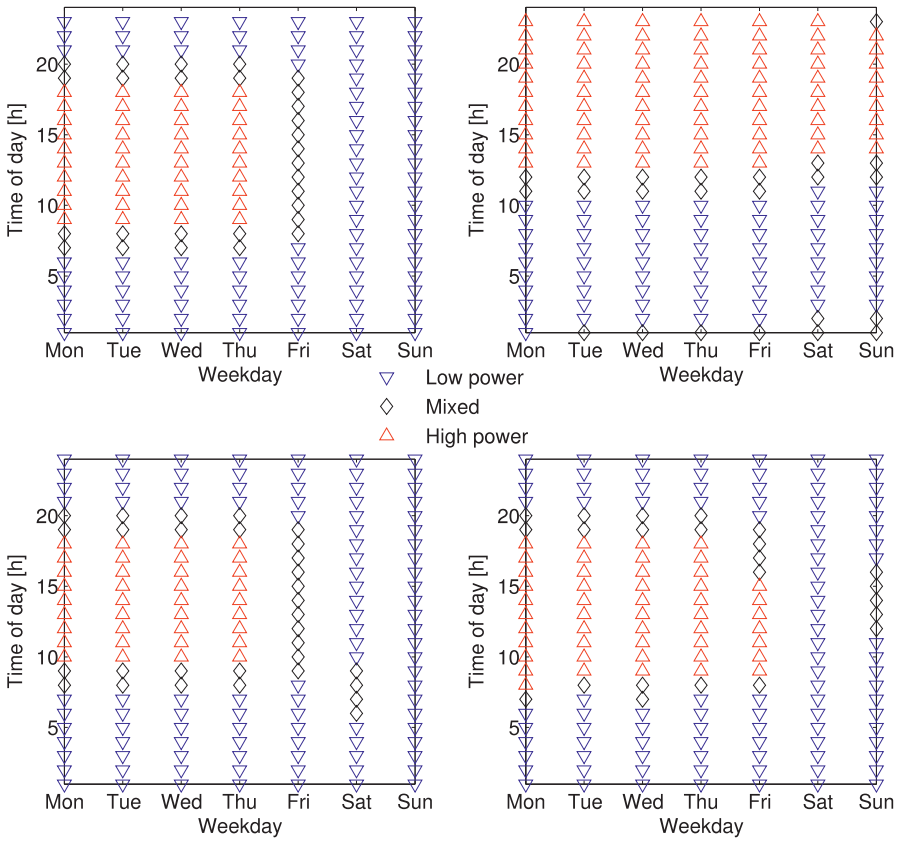


Figure 3.10. Weekly schedules of three substations with high BC (substations #1-4 in Figure 3.8). These schedules were automatically generated using the function `power_schedule` that is included in the Appendix. The corresponding power profiles are displayed in Figure 3.11.

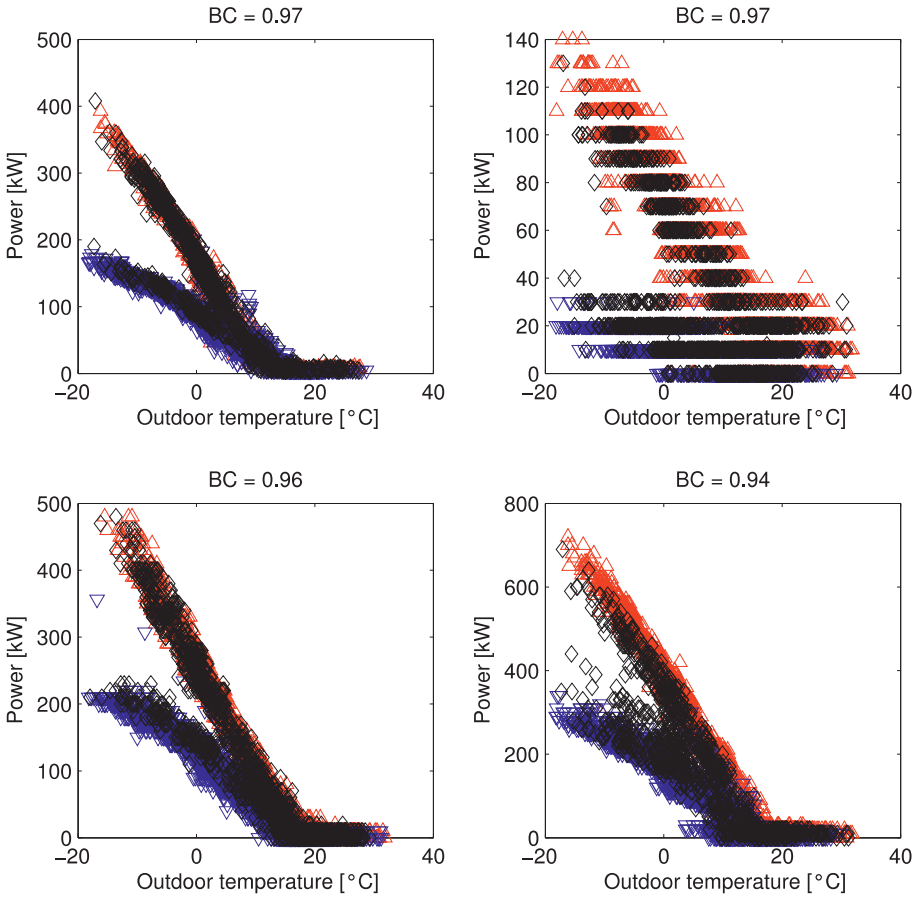


Figure 3.11. Power profiles of four substations with high BC (substations #1-4 in Figure 3.8). Symbols indicate the clusters defined by the weekly schedules illustrated in Figure 3.10 and denote high power (triangles, pointing upwards), low power (triangles, pointing downwards) and mixed power (diamonds).

The separation of different power clusters can be further improved by considering additional features and using linear discriminant analysis to transform the features into a new basis where eventual clusters are maximally separated, see Li et al. (2010) for an example based on daily average power and canonical discriminant analysis. However, the aim here is to detect major cycles in the power, which motivates *automated* calculation of a weekly schedule and multiple regression models of different branches in the power profile. Therefore, detection of minor cycles and level differences in the standardized power is not desirable here, and a simple approach based on the standardized power and k-means analysis is used.

There will be exceptions to a weekly schedule of this type, in particular at public holidays and in the event that the thermal load pattern changes. This will result in a higher variance of the regression model residuals, and outliers that may be automatically identified as anomalies if the change is sufficiently large. Therefore,

data at public holidays needs to be handled separately. Li et al. (2010) handles all public holidays like Sundays. Seem (2005; 2007) categorizes holidays in the same way as other days using the average and peak daily power, which automatically leads to a proper categorization of holidays. The weekly schedule is a trade-off between well-defined system behaviour and flexibility, which enables an operator to interpret the result and re-define or re-calculate schedules when needed.

Next we describe how a regression model can be fitted to the power profile, once the eventual cycles in the power have been identified and quantified in terms of a weekly schedule as described above. In this report we use piecewise linear regression to model the relationship between the power and the outdoor temperature, but also other relationships between quantities. Therefore, the discussion of linear regression is factored out to a separate section of this chapter. Note that power data that belongs to the mixed class is not modelled explicitly because it should comply either with the low-power model, or the high-power model.

3.2 Piecewise linear regression

Linear regression is a probabilistic approach to modelling the relationship between a dependent variable, y, and an explanatory variable (or vector), x, which has many practical uses. It is extensively used in practical applications because models that depend linearly on the parameters are easy to fit and the properties of the resulting estimators are easier to determine compared to models that are non-linear in the parameters. Multivariate or multiple linear regressions are used when the explanatory variable is vector valued. Linear regression is a standard tool in statistics and machine learning, which is used both for modelling relationships between variables and for classification. See Murphy (2012), Chapter 7 for a formal introduction to linear regression. Here the goal is to predict the expectation value of the dependent variable (for example the power) given an explanatory variable (for example the outdoor temperature) using a *piecewise linear regression model*, which is motivated by the discussion in the beginning of this chapter concerning the relationship between the power used for space heating and the outdoor temperature.

3.2.1 Breakpoints

Piecewise linear regression can be used to fit multiple linear models to data for different ranges of the independent variable, x; see Figure 3.1 for an example. The values of the independent variable where the slope of the linear function changes are called *breakpoints*. The optimal positions of breakpoints are problem-specific and may or may not be known in advance. If the number and positions of breakpoints are unknown they can be estimated with a data-driven optimization process or heuristics. The piecewise regression function can be made continuous at the breakpoints, which is the natural choice here because we expect that the average power should vary continuously with the outdoor temperature. In order to illustrate the approach we

consider an example with only one breakpoint located at $x = c$. A piecewise linear model can then be written on the form

$$P = a_1 + b_1 x \text{ for } x \leq c, \tag{11}$$

$$P = a_2 + b_2 x \text{ for } x > c. \tag{12}$$

Continuity at the breakpoint, $x = c$, requires that

$$a_1 + b_1 c = a_2 + b_2 c \rightarrow a_2 = a_1 + c\ b_1 - b_2\ . \tag{13}$$

Therefore, the continuous piecewise linear model can be written on the form

$$P = a_1 + b_1 x \text{ for } x \leq c, \tag{14}$$

$$P = a_1 + c b_1 + b_2 (x - c) \text{ for } x > c. \tag{15}$$

This approach can be generalized to an arbitrary number of breakpoints. If the positions of the breakpoints are known the resulting system of equations can be solved directly in a least-squares sense, otherwise an optimization procedure / non-linear least squares must be used, which is called adaptive linear regression. There are several methods for adaptive linear regression, which are based on different assumptions about the optimal way to select the number and positions of breakpoints. We list some examples of software packages that can be used for adaptive and non-adaptive linear regression in the Appendix. Optimization of breakpoints is typically done in two steps, first by adding breakpoints to reduce the variance, then by pruning the breakpoints with a penalty function so that a reasonable trade-off between accurate fit and model complexity is achieved. The purpose of that procedure is to avoid over-/under-fitting, but it is tricky to implement in a reliable way for automated generation of models. Our experiments suggest that tuning of the adaptation parameters and assumptions are needed in order to avoid over fitting with adaptive methods. This is not problematic when using adaptive regression in a manual fashion, but it is challenging in an automated setup where regression models are calculated automatically for thousands of substations without human assistance. Therefore, we do not use optimization procedures, but a fixed number of pre-defined breakpoints.

3.2.2 Robust regression

More important than adaptation of breakpoints is that the linear regression method is robust to outliers, otherwise faults in the historical record can bias or invalidate the regression models that are used as reference models for fault detection. Robust linear regression is supported by several software packages that are listed in the Appendix. See Murphy (2012), Section 7.4-7.5 and Table 8.1 for an introduction to robust regression. The basic idea is to assume a heavy-tailed probability density function for the data when fitting the regression model, which implies that the effect of a small

subset of outliers on the expectation value is negligible. Another principle used to reduce the impact of outliers, in particular when fitting polynomials of higher order (like cubic splines) with linear regression is to introduce a penalty function that favours polynomials with small coefficients. Note that linear regression is not limited to linear functions; it is possible to fit polynomials of arbitrary order and other non-linear functions of the data using linear regression, provided that the unknown coefficients are linear functions.

3.2.3 Test results

We use the Splinefit tool to automatically fit piecewise linear regression models to the 996 substations in the test data set; see the Appendix for further information. The particular regression algorithm used is not critical, but it should be a robust regression algorithm. We identify a subset of 853 substations with a BC below 0.6, indicating that there are no significant intraday or intraweek cycles in the power, and we automatically fit piecewise linear regression models to the annual power profiles of these substations. A subset of substations with low BC is used here to simplify and speed up the analysis by avoiding the cluster analysis step. In the following we present a residual analysis of the piecewise linear models for different assumptions about the breakpoints. We quantify the residuals in terms of the root-mean square (RMS) error, ϵ_{RMS}, of the residuals relative to the expectation value, $E[y_i]$, given by the regression model,

$$\epsilon_{RMS} = RMS\ \frac{y_i - E[y_i]}{E[y_i]}\ . \tag{16}$$

The number and position of breakpoints can be defined in different ways. A basic assumption is that there is only one breakpoint located at the onset of space heating, which typically is 2-3°C below room temperature. For example, with a single break point at 16°C the RMS error for the population of 853 substations is 4.3. The high RMS error is a consequence of the relatively high standard deviation at high outdoor temperature compared to the expectation value, $E[y_i]$; see Figure 3.1 for and example. In addition, we use a random 50% subset of the annual data for each substation to fit the regression model, and the remaining 50% of the data to calculate the regression model residuals. Therefore, the RMS error is an estimate of the *generalization* error of the regression model.

Additional breakpoints can be defined manually, or by dividing the outdoor temperature interval into subintervals of equal length, or into subintervals defined by an approximate equal number of observations within each interval. The first two approaches require application-dependent knowledge about the climate and substation data, so that overfitting resulting from too few data points in some interval can be avoided, in particular at low outdoor temperatures where the data can be sparse. The latter approach is straightforward to automate because it does not depend on the range or distribution of the dependent variable, and it ensures that the numbers of data

points within the intervals are comparable. Therefore, we use the second, frequency-based positioning approach in this work. The breakpoints can either be defined individually for each substation, or they can be determined jointly for the whole population of substations so that all regression models share the same set of breakpoints. The former approach result in lower RMS error and is preferred, unless there is a need to have identical breakpoints for all substations, for example to simplify visual comparison of regression models.

We calculate the annual RMS error for the 853 substations for a varying number of breakpoints, which are defined independently for each substation so that the number of data points that falls in-between each adjacent pair of break points are equal on an annual basis (approximately equal when the number of data points divided by the number of intervals is not an integer). We use a random 50% subset of the annual data for each of the 853 substations to fit the regression models, and the remaining 50% of the data to calculate the regression model residuals. As we describe above the purpose of this procedure is to estimate the generalization error of the piecewise linear regression model. The result of this calculation is presented in Table 3.1.

Table 3.1. Generalization RMS error versus the number of breakpoints and piecewise linear segments used to model the relationship between the outdoor temperature and the thermal power for 853 substations with $BC < 0.6$.

Number of breakpoints	Number of segments	RMS error
0	1	172
1	2	125
2	3	56
3	4	8.0
4	5	3.6
5	6	2.6
6	7	2.5
7	8	2.4
8	9	2.5
9	10	2.5

According to the table above the RMS error with four frequency-based breakpoints is 3.6, which is slightly lower that the RMS error obtained with one breakpoint at 16°C. By introducing up to seven breakpoints and eight segments the RMS error can be further reduced, which indicates that a maximum of eight segments should be used. There is no benefit of adding more breakpoints in terms of variance because a higher number of breakpoints result in a slightly higher RMS error, which may indicate over-fitting. Six regression models that are fitted to the power profile of an apartment building with a BC of 0.6 are illustrated in Figure 3.12, for a varying number of breakpoints.

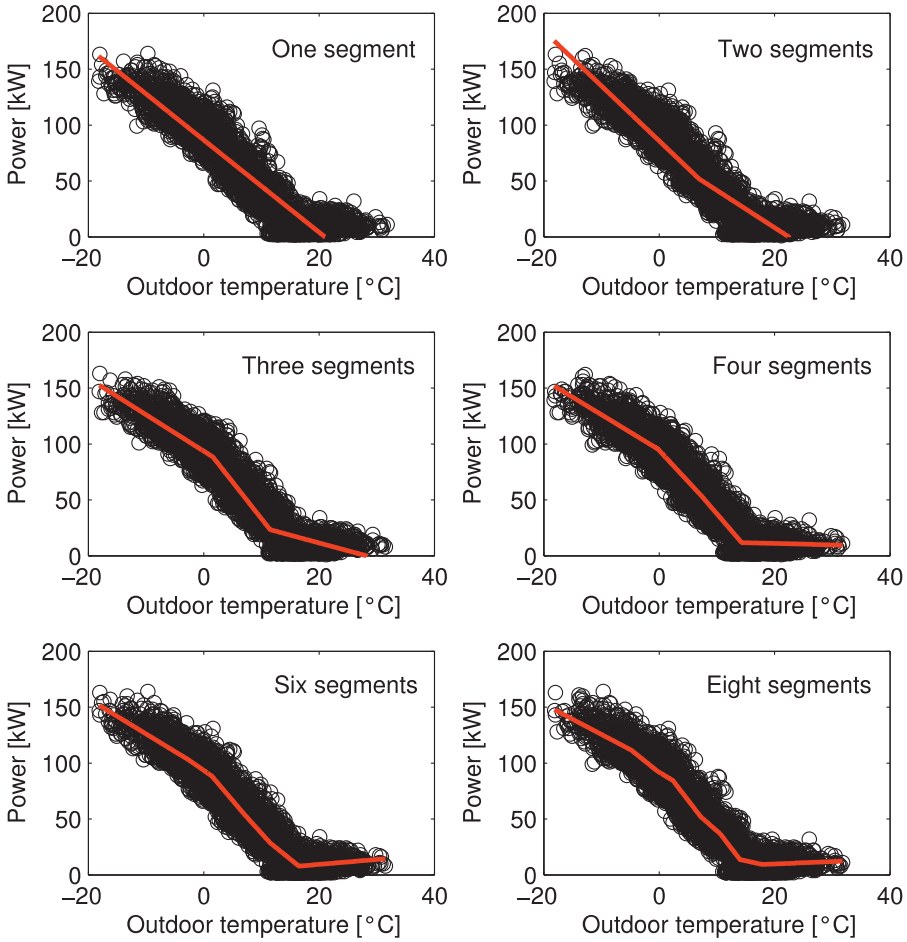


Figure 3.12. Six piecewise linear models that are fitted to the power profile of an apartment building. The BC for this building is 0.6 and there are no significant time-varying cycles in the power. The six models are fitted with a varying number of breakpoints and segments, as indicated by the labels within the panels.

For substations with cycles in the power the weekly schedule resulting from the cluster analysis is used to divide the data into low- and high-power clusters, and a piecewise regression model is fitted to each cluster; see Figure 3.13 for an example. This substation has evident intraday and intraweek cycles in the power and a BC of 0.97; see Figure 3.5.

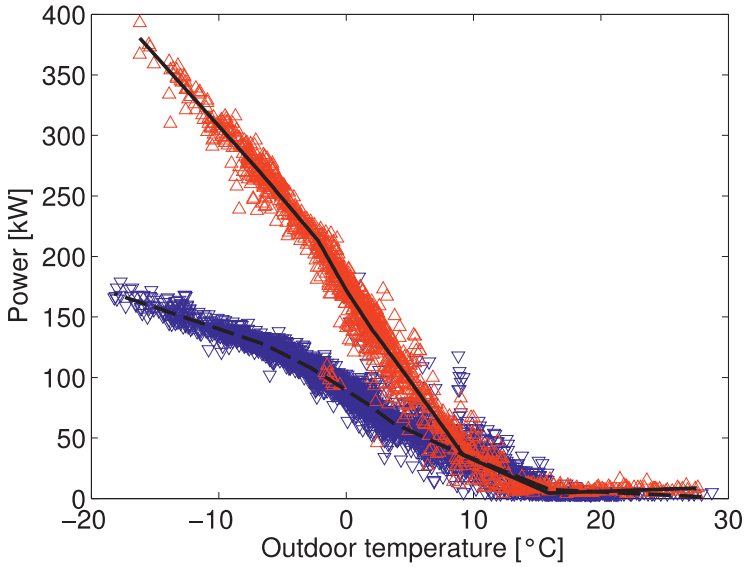


Figure 3.13. Two piecewise linear models that are fitted to the power profile of an office building in Stockholm with two branches in the power profile. This building has a BC of 0.97. The two regression models are fitted automatically using a weekly schedule to group the data into high- and low-power categories.

Some misclassified data points are visible in Figure 3.13. Some of these data appear at public holidays. For example, the small group of misclassified data points at about -2°C appears at the end of December and beginning of January.

3.2.4 Residual analysis

The residuals of a fitted model are the differences between the observed values and the prediction calculated using the regression model, $\varepsilon_i = y_i - E[y_i]$. If the model fits the data well the residuals should be randomly distributed. If the residuals have an evident non-random structure the model fits the data poorly. The purpose of the discussion above about intraday and intraweek cycles in the power, and how to analyse and describe such cycles with cluster analysis and a weekly schedule, is to remove evident cycles in the residuals of the regression model of some substations. By fitting separate regression models to the high- and low power categories the variance of the residuals are reduced significantly. For substations with a low BC and only one cluster a single regression model is used. We select four random substations from the population of 996 substations and calculate normal probability plots for

these four substations, see Figure 3.14. A normal probability plot illustrates the probability of deviations from the expectation value, both for the data sample and an idealized normal distribution. Another common plot that is used to illustrate the distribution function of variables is the QQ-plot (Murphy, 2012, Section 8.4.5), which displays the quantiles of the data versus the quantiles of a reference distribution.

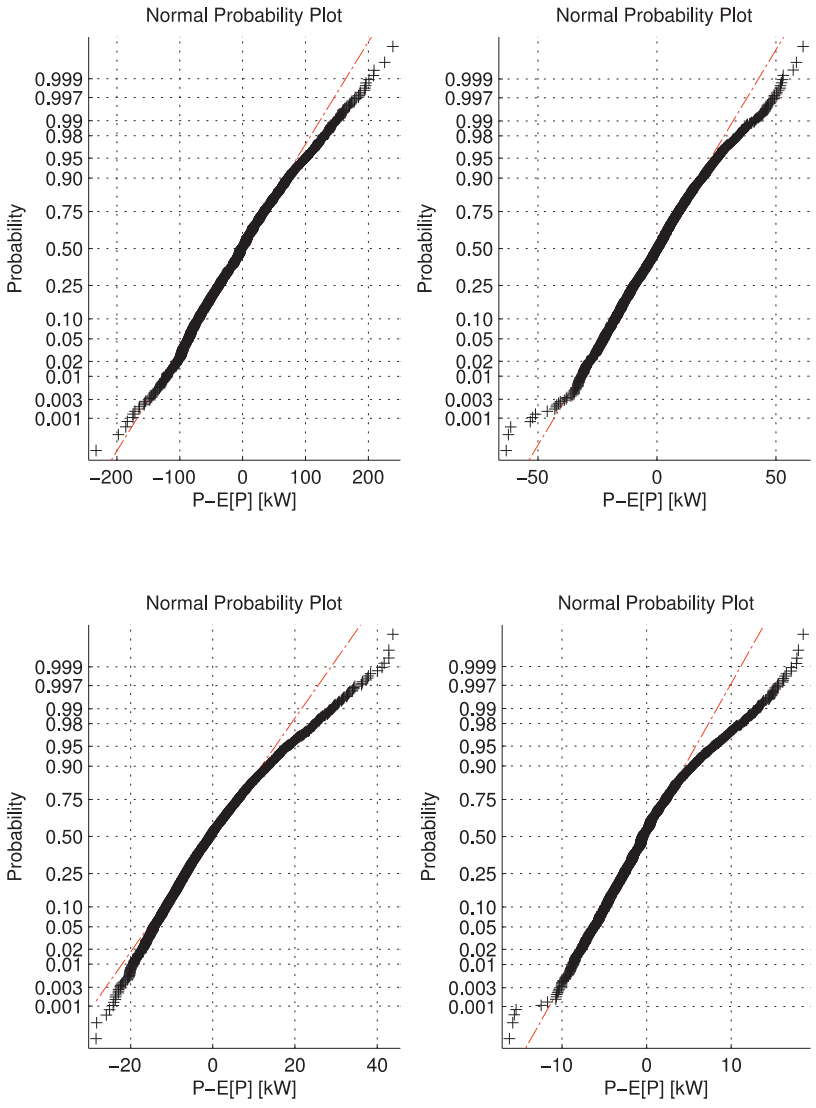


Figure 3.14. Normal probability plots of the regression-model residuals for four different substations that are selected randomly from the population of 996 substations. The dash-dotted line represents a normal distribution. In all four cases the upper tail is short, while the lower tail is either short or too heavy.

The probability distribution functions of the model residuals vary between substations, and they are often skew and rarely normally distributed. The deviation

from normality is not surprising because there are several hidden variables and states that affect the power used by a district heating substation that are not accounted for by the piecewise linear relationship between the outdoor temperature and power (formally, the central-limit theorem is not applicable here because of the character of the variables involved and the limited size of the one-hour samples). For example, the thermal dynamics of the building, effects of sun and wind, details in the control of the ventilation system and human behaviour can affect the thermal power. The linear regression approach proposed here is still useful because a large sample is used to generate the model, and a simple model and fitting procedure is motivated by the need to automatically calculate thousands of regression models for substations that have complex and varying characteristics. It is possible that a more accurate model can be defined by integrating a dynamic model of the building with a probabilistic model that describes the cycles, but we have not investigated that possibility. Our empirical results show that the piecewise linear model is sufficiently accurate to be useful for the identification of outstanding anomalies and faults, and it gives sensible results when applied to a sizable population of substations in an automated fashion.

3.2.5 Imputation of missing data

Imputation is the problem of estimating missing data. Faults in the communication with energy meters, or energy meters that are rebooted or run out of power are common causes of missing energy metering data. Linear interpolation is commonly used today for the estimation of missing energy data, which is a crude estimate when the outdoor temperature varies or there are time-dependent cycles in the power. A regression model of the type introduced above provides a more accurate method for imputation of missing energy data, which accounts for both the temperature-dependence and eventual time-dependent cycles in the power.

3.2.6 Implementation

Several processing steps are required to calculate regression models of the power profile. A block diagram of the key steps is illustrated in Figure 3.15. The first step is to extract features from the energy meter data. The BC is calculated and a decision is made whether a weekly schedule should be created. The decision to proceed with the cluster analysis is based on the value of the BC according to the discussion above. Alternatively, the cluster analysis is performed for all substations and the result is used to determine whether a schedule needs to be created. The latter approach is more reliable because the BC can be low when there are cycles in the power (we have not observed that but in principle it may happen), but also requires more computational resources. If clusters are detected the data is grouped into high-, mixed- and low-power categories, and regression models are fitted to the high- and low-power categories. Only one regression model is fitted if there are no evident cycles / clusters in the power profile.

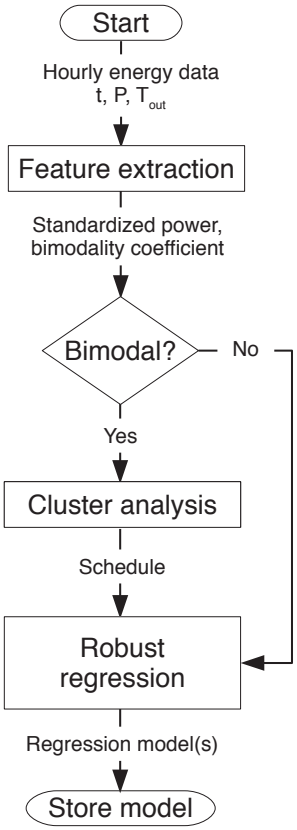


Figure 3.15. A block diagram illustrating the computational steps needed to determine the regression model(s) of a substation power profile.

Regression models can be fitted to data received during a specific interval of time, which we refer to as a *reference period*. By comparing regression models that have been fitted at different reference periods it is possible to detect and illustrate changes in the power. This is useful for fault diagnosis and for the communication with customers when a fault is detected and bills have to be adjusted (Reference group). By plotting the power data and regression models before and after the introduction of the fault a major change becomes evident.

The bimodality coefficient and weekly schedule (when applicable) can be automatically calculated for the whole population of substations for a reference period, preferably at low outdoor temperature. An interface that enables management of expected deviations from the schedules is needed, in particular for the management of holidays. Schedules should be calculated for a reference period, and it should be

possible to create different schedules for different periods of time because the cycles can change over time.

If the historical data that is used to calculate the bimodality coefficient and weekly schedule is faulty the result can be incorrect. Therefore, it is important that this tool is combined with tools for anomaly detection and ranking, so that incorrect schedules can be detected and recalculated using data from another reference period. Methods that can be used to detect such anomalies are discussed in the next two chapters of this report. This approach can also be used to model the flow and the supply temperature, but no cluster analysis is needed in the case of the supply temperature. We discuss these points below.

3.3 Flow

The primary flow can be modelled much like the power because the power is controlled with the flow valve(s). This means that the block diagram that is illustrated in Figure 3.15 can be applied to calculate regression models of the flow also. The methods for anomaly detection that we discuss in this report do not require that a regression model of the flow is used, but such models can be helpful for diagnosis purposes when an anomaly is detected and a manual investigation of the substation data is required. An example of the relationship between the outdoor temperature and flow is illustrated in Figure 3.16, which includes also a piecewise linear model that is fitted to the flow data.

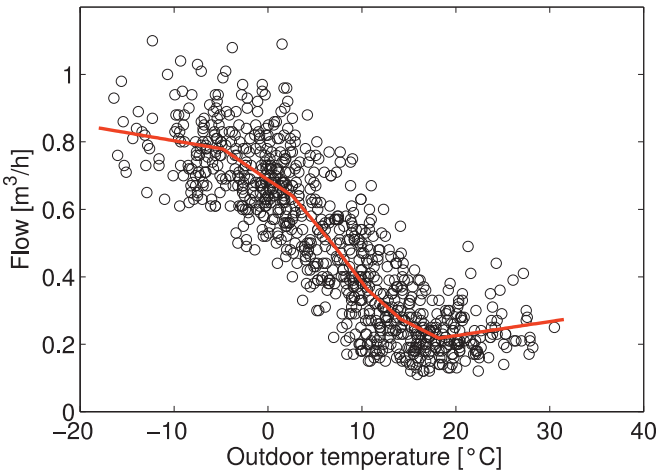


Figure 3.16. The primary flow versus the outdoor temperature for an apartment building. The power profile of this substation is illustrated in Figure 3.1. The solid line is a piecewise linear regression model that I fitted to one year of data, but only 10% of the data is displayed (circles).

A regression model of this type can be used for diagnosis purposes to detect outliers in the flow with the method that is discussed in Chapter 4, signal drift with the

method that is described in Chapter 5, and precision problems or abnormal noise levels with the method that is described in Chapter 6.

A well-known problem is that the primary supply can be manually short-circuited during the summer and that an open valve is forgotten, which results in abnormal flow and inefficient operation of the substation. This problem can be detected with regression modelling and outlier detection.

3.4 Primary supply temperature

In Chapter 2 we discuss two basic methods for fault detection using the primary supply temperature. Limit checking with a constant threshold that is set slightly above the maximum expected temperature is useful for detection of large deviations, for example as a result of faults in the cabling or electronics. The second basic test is that the primary return temperature should be lower than the primary supply temperature.

It is possible to improve on these basic limit-checking approaches by modelling the relationship between the outdoor temperature and the primary supply temperature with a piecewise linear regression model; see Figure 3.17 for an example.

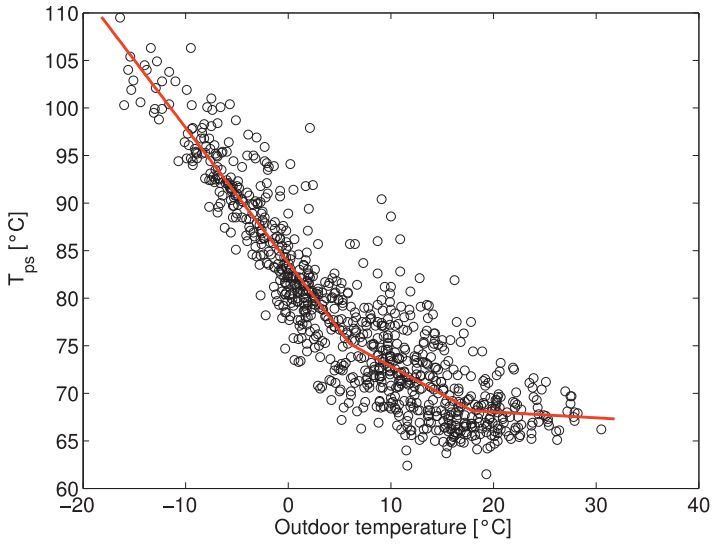


Figure 3.17. Primary supply temperature versus the outdoor temperature. This is a typical example that illustrates how the supply temperature of a substation can be modeled in terms of the outdoor temperature. The regression model (solid line) has four breakpoints and it is fitted to one year of hourly data. For clarity, only 10% of the data (circles) are illustrated in the figure.

The primary supply temperature varies with the outdoor temperature because the temperature that is supplied to the network from the heat production plant is controlled with weather forecasts. Using a regression model of the supply temperature it is possible to apply the outlier detection and ranking method that is described in

Chapter 4, to detect signal drift with the method that is described in Chapter 5, and precision problems or abnormal noise levels with the method that is described in Chapter 6. Therefore, the implementation and algorithms needed for fault detection with power data can be reused with minimal changes for fault detection with the supply temperature.

3.4.1 Comparing neighbours in the network

An alternative approach is to make use of the redundancy in the supply temperatures measured by different substations, which are connected to the same network at locations where the supply temperatures are similar (Sandin et al., 2012). For example, substations that are connected at nearby positions along the same supply pipe may have similar supply temperatures, provided that the flow is reasonably high so that cooling in the connection pipe is low, see Figure 3.18.

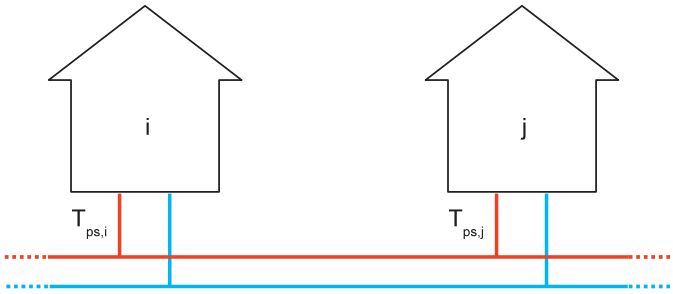


Figure 3.18. The primary supply temperature, $T_{ps,i}(t)$, of one substation, i, can be compared with the supply temperature, $T_{ps,j}(t)$, measured at another substation, j, located in the same network at a position where the supply temperature is similar.

The supply temperature varies with geographical location because there are thermal losses within the distribution network, and this effect is more prominent at long distances and low flow. Therefore, it is important to compare substations that have *similar* supply temperatures. If the network structure and coordinates of the substations are known that information can be used to identify the substations that should be compared. Another approach, which is straightforward to automate and does not require knowledge about the network structure and geographical coordinates, is to analyse the *correlation* between supply temperature time series, $T_{ps,i}(t)$ and $T_{ps,j}(t)$. The correlation coefficient is defined in terms of the covariance matrix of the two time series. In general, the correlation coefficient $\rho_{x,y}$ of two variables x and y is defined as

$$\rho_{x,y} = \frac{cov(x,y)}{\sigma_x \sigma_y} = \frac{E[\ x-\mu_x \quad y-\mu_y\ ]}{\sigma_x \sigma_y}, \tag{17}$$

where μ denotes the average and σ is the standard deviation.

In order to identify substations with similar supply temperatures using correlation analysis it is necessary to de-trend the time series data, otherwise it is difficult to discriminate between the substation pairs because the covariance is dominated by the seasonal trend in the supply temperature. Common de-trending approaches include (Box et al., 2008): differencing, curve fitting, filtering, and piecewise approximation with polynomials. There are two options that are straightforward to implement here; either the expectation value given by a piecewise regression model is used to subtract the seasonal trend, or a differencing / high-pass filtering approach is used. A time series that has a non-stationary average (seasonal trend) can be made stationary by taking the first-order difference of the samples, x_i, at position i and $i-1$; $w_i = x_i - x_{i-1}$. First-order differencing is used here because it is straightforward to implement and it does not require a long-term record of data, which would be the case if a regression model were used. This approach is implemented in the function named `find_correlated` in the Appendix.

3.4.2 Test results

We calculate the correlation coefficient between de-trended primary supply temperatures of each pair of substations in the test set, $\rho_{dT_{ps,i},dT_{ps,j}}$, where $dT_{ps,i}$ denotes the first-order finite differences of the supply temperature time series. Figure 3.19 illustrates the correlation coefficient between the de-trended supply temperature to one apartment building and the de-trended supply temperatures to the other 995 substations in the test set. A subset of the substations has highly correlated supply temperatures because the correlation coefficients are close to one.

For each substation in the population we calculate the correlation coefficients with respect to all other substations, and the maximum correlation coefficient is identified in each case. In this way the pairs of substations that have maximally correlated (de-trended) supply temperatures are identified and can be used for the subsequent comparison of supply temperatures, forming the basis for fault detection. We calculate the maximal correlation coefficients and the corresponding geographical distances between substations for each substation in the test set using one year of supply temperature data, see Figure 3.20.

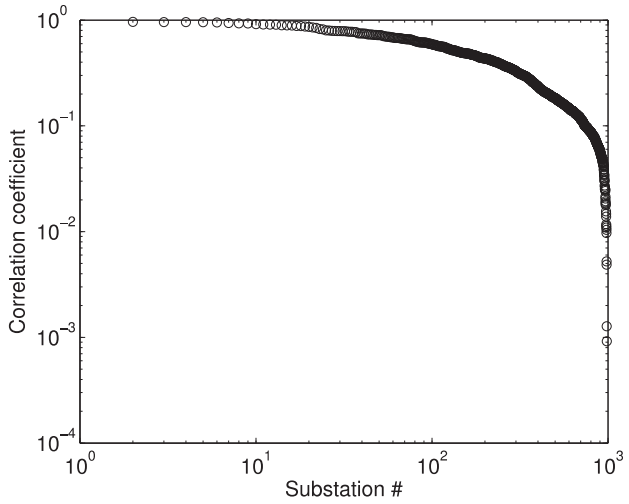


Figure 3.19. The correlation coefficients between primary supply temperatures of one apartment building compared to all other substations in the test set. The supply temperature to this building is illustrated in Figure 3.17.

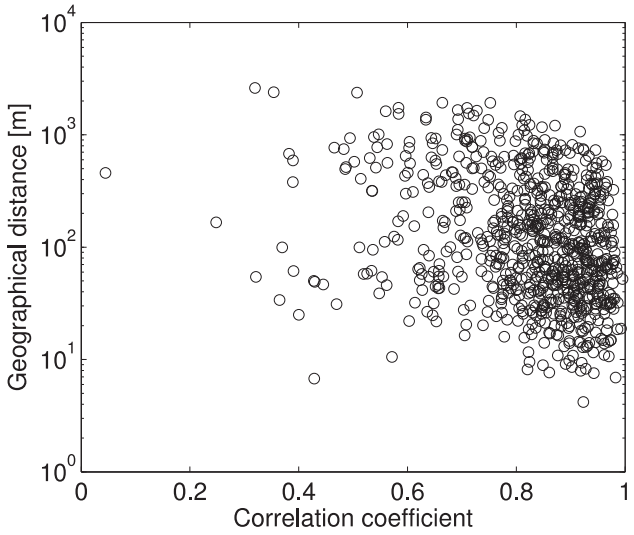


Figure 3.20. Maximum correlation coefficients for the whole population of substations in the test set versus the corresponding geographical distances.

This result shows that for most substations in the test set there are other substations that have similar supply temperatures in terms of high correlation coefficients. A few substations have low maximum correlation coefficients in comparison to the remaining population, which indicates that these substations have abnormal supply temperatures. However, this analysis is limited to 996 substations, which is a fraction

of the substations in the whole network. Therefore, low correlation coefficients can result also from the limited data set. In a full-scale analysis, where all substations of the network are included, an exceptionally low correlation coefficient motivates further diagnosis.

There is no evident relationship between the maximum correlation coefficients and the corresponding geographical distances between substations, which indicates that the geographical distance alone is insufficient for the identification of substations with similar supply temperatures. The identification of pairs of substations with similar supply temperatures using the correlation coefficient can possibly be useful also in networks where substations are allowed to contribute power to the network, for example in the form of excess heat from industrial processes. In that case an analysis based on network structure may be difficult, but the correlation analysis approach presented here remains simple and can still be automated.

The difference of supply temperatures measured by two maximally correlated substations is defined as

$$\Delta T_{ps}(t) = T_{ps,i}(t) - T_{ps,j}(t), \tag{18}$$

where t is the time and i, j is a pair of substations with maximally correlated detrended supply temperatures. The supply temperature difference, ΔT_{ps}, is expected to have a low variance when the flow is high in both substations, while low flows may be associated with higher variance because of cooling in the pipes. Therefore, we consider the relationship between the supply temperature difference and the geometric mean of the flow, which is the square root of the product of the two flows. A geometric mean is used because it normalizes the ranges of the variables being averaged, so that no variable dominates the weighting and a high flow in one substation can compensate for vanishing flow in the other substation. Figure 3.21 illustrates an example of the supply temperature difference for the substation that is illustrated in Figure 3.17. The maximum correlation coefficient is 0.84 for this substation. A piecewise linear regression model that can be used for anomaly detection (Chapters 4-6) is fitted to the data.

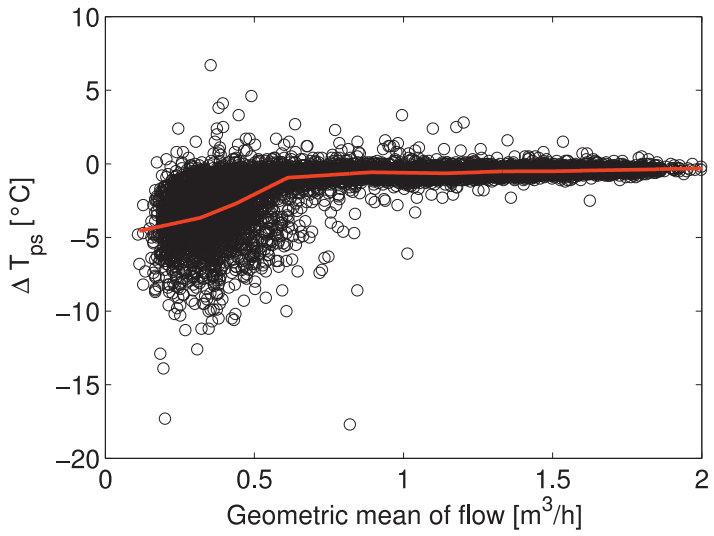


Figure 3.21. Difference of primary supply temperatures measured by two different substations with maximally correlated supply temperatures versus the geometric mean of the flow. A piecewise linear regression model with eight segments is fitted to the data, which can be used for fault detection.

This is a typical result, where the variance is high at low flow and vice versa, which is a consequence of cooling in the pipes. Note that there can be an offset in the supply temperature difference because the temperature sensors of flow meters are often calibrated in pairs, so that the difference between primary supply and return temperatures is measured accurately, but not necessarily the absolute temperatures. Therefore, the offset from $\Delta T_{ps} = 0$ in the regression model is arbitrary and is not considered as an indicator of anomalies. The magnitude of the temperature difference illustrated in Figure 3.21 is smaller than that illustrated in Figure 3.17, which means that this is a more accurate model of the supply temperature than a model that is fitted directly to ΔT_{ps} T_{out} , in particular at low outdoor temperatures when the flow is high.

3.5 Return temperature

The primary return temperature is more difficult to model compared to the other variables (power, flow and supply temperature) because several hidden processes affect it, in particular the use of heated tap water. When working with hourly data the return (and supply) temperature is sampled once per hour, which means that a short-term fluctuation in the return temperature at the time when the sample is recorded will result in an abnormal deviation of the hourly value. Therefore, the variance in the return temperature is significantly higher than the variance of the supply temperature; see Figure 3.22 for an example, and Figure 3.17 for the corresponding primary supply temperature. The relationship between the average return temperature and the outdoor

temperature is complex and varies significantly between substations. These circumstances make modelling and anomaly detection beyond the level of basic limit checking difficult.

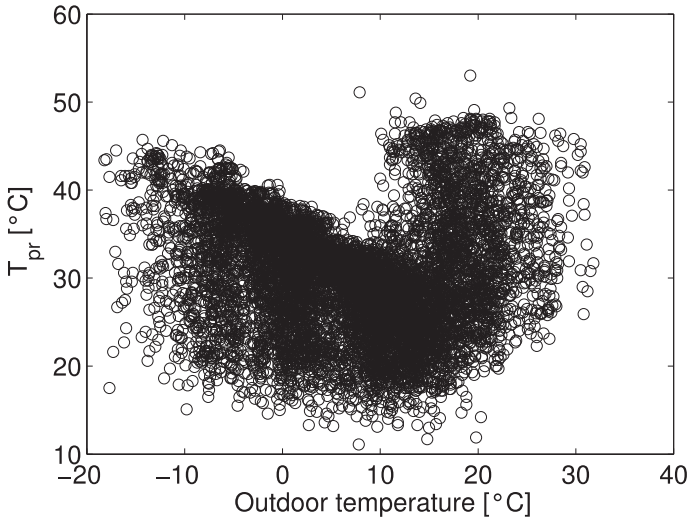


Figure 3.22. The primary return temperature versus the outdoor temperature.

We have not identified an accurate method for anomaly detection of the primary return temperature. Unlike the case with the primary supply temperature there is no explicit redundancy in the data that can be used to validate the return temperature. A relatively simple modification that would reduce the variance of the return temperature is to calculate the hourly value with a digital low-pass filter within the energy meter, instead of sampling a momentaneous value. See Bergquist et al. (2004) for a discussion about low-pass filtering of variables in district heating substations.

4 OUTLIER DETECTION

An *outlier* is an observation that appears to deviate markedly from other members of the sample in which it occurs. Typically this means that the numerical value of an outlier observation is distant from the rest of the data, see Figure 2.4 for an example. Data sets may also contain multiple outliers. Outliers can occur by chance when sampling a random variable, but they can also indicate novel observations in an experiment that does not fit the model or expectation (cf. black swan metaphor) and they can result from measurement errors. Here we are mainly interested in the latter aspect, where outliers may indicate that measurements are faulty.

A common cause of outliers is that values are sampled from a mixture of multiple distributions, which in practice may not be fully understood and described. Such distributions can be modelled with mixture models, or the sub-population identities of the observations can be identified with clustering or unsupervised learning methods. Here we use the latter approach. The bimodality coefficient and weekly schedule of load cycles that are introduced above are used to identify whether values correspond to high- or low thermal load distributions before we test for eventual outliers in the data. This approach is transparent and reduces the variance of the model residuals significantly for substations with cycles in the load.

Outliers play an important role also when calculating statistical models of data. In that context *robust* estimators that are insensitive to outliers are used. An outlier *score* can be used to quantify by how much an outlier deviates from the expectation. The Box plot (Tukey, 1977) is a commonly used graphical tool for the identification and illustration of outliers, see Seem 2007 for an example. In this chapter we describe a method that can be used for automated detection and scoring of outliers, which can be used to identify outstanding outliers in energy metering data of a large population of substations.

4.1 GESD test for outliers

Several tests for outliers exist and there is no generic best choice. However, the generalized extreme studentized deviate (GESD) test (Rosner, 1983) is recommended when the number of outliers is unknown because it works well under a variety of conditions (Iglewicz & Hoaglin, 1993). In particular, the GESD method has been studied and proposed for the detection of abnormal energy use in buildings (Seem, 2007; Li et al., 2010). Therefore, we introduce this algorithm here and provide an implementation in the Appendix. Further information and examples are available online (NIST, 2012, Section 1.3.5.17) and a detailed step-by-step description of the algorithm can be found in Seem, 2007.

Given an upper bound on the number of potential outliers, r, the GESD method is defined for a hypothesis test of the following type.

Hypothesis	Description
H_0	There are no outliers in the data set.
H_i	There are i outliers in the data set.

The test statistic is iteratively defined in terms of the standardized maximum absolute deviation of the sample

$$R_i = \frac{\max_j x_{i,j} - E[x_i]}{\sigma_{x_i}}, \quad x_1 = x, \quad x_i \setminus x_{i,j} \to x_{i+1}, \tag{19}$$

where $E[x]$ is the expectation value or mean of the sample and σ_x is the standard deviation. In this work we calculate the expectation value using piecewise regression models that are fitted to data from a historical reference period (one exception occurs in Chapter 2, where we use a simplified approach based on a weekly moving average for the estimation of the mean power). When the first statistic, R_1, has been calculated from the n observations in the sample $x_1 = x$ the value $x_{1,j}$ that maximizes the absolute deviation is removed from the sample. This step is defined by the set operation $x_i \setminus x_{i,j} \to x_{i+1}$, which implies that the maximum deviation $x_{i,j}$ is removed from the set x_i. The subsequent R_i are calculated from the remaining values in x. This procedure is repeated until r extreme values have been removed from the sample x and the resulting standardized statistics $R_1, R_2, R_3, \ldots, R_r$ are known.
For each statistic, R_i, the following test quantity is calculated (Rosner, 1983)

$$\lambda_i = \frac{(n-i)\, t_{p,n-i-1}}{\sqrt{(n-i-1+t_{p,n-i-1}^2)(n-i+1)}}, \tag{20}$$

where $t_{p,\nu}$ is the inverse of the cumulative Student's t-distribution with ν degrees of freedom and a tail area probability p that is defined by

$$p = 1 - \frac{\alpha}{2(n-i+1)}. \tag{21}$$

Here α is the (idealized) significance level, for example $\alpha = 0.05$ for a 95% confidence level. The cumulative Student's t-distribution is a standard function that can be calculated with most statistical software packages. This function can also be expressed in a regularized incomplete beta function (Olver et al., 2010, Section 8.17). See the implementation of the GESD outlier test in the Appendix for further information. The number of outliers in the sample is given by the maximum i so that

$$R_i > \lambda_i. \tag{22}$$

Simulation studies indicate that this test is accurate for $n \geq 25$ and reasonably accurate for $n \geq 15$ (Rosner, 1983), which is easily achieved when dealing with hourly energy metering data. Formally, these results are valid only when the probability density function of the sample is approximately normal because the test

statistic involves only the mean and standard deviation of the sample. Nevertheless, our empirical studies indicate that the GESD method is useful for the detection of outliers in energy metering data, which is in line with the conclusions by others (Seem 2007; Li et al., 2010). The GESD method is also straightforward to implement in automated fashion, which is an important aspect if the method is to be implemented and used by the industry.

4.2 Ranking of outliers with Z scores

Outliers can be scored depending on how much they deviate from the expected value. Scoring is a helpful tool for ranking of outliers, for example when a high number or high rate of outliers in the data prevents manual inspection of all potential outliers. In the context of hourly district energy data, ranking of outliers is necessary because deviations from ideal behaviour are expected for natural reasons (energy meters have limited precision, the one-hour sampling interval results in inconsistencies between the variables etc., see Chapter 1). Therefore, methods for ranking of anomalies are needed to enable fast detection of outstanding outliers and investigation of potential faults. Ranking methods are also useful for the detection of anomalies that are not related to faults in the substation instrumentation, but which may be of interest for the customers. For example, a problem with the ventilation that affects the load cycles can be detected and provided as an information service to the customer.

A basic approach is to score the outliers with the number of standard deviations of the outlier value, a so-called standard score or *Z score*

$$z_i = \frac{x_i - E[x]}{\sigma_x}. \tag{23}$$

Here x_i is the value of an observation that has been identified as an outlier with the GESD method and z_i is the Z score of the outlier. For example, a Z score of 10 implies that the outlier deviates from the expectation value by ten standard deviations. In addition to the magnitude of the Z score, the sign of a score is informative because it indicates whether the value of an outlier is higher or lower than the expected value.

A Z score that is defined in this way is not robust. If there are outliers in the sample the standard deviation can be overestimated, which results in low Z scores and an underestimate of the potential importance of the outliers. Also, the mean is not a robust statistic. Therefore, it is common practise to use a *modified Z score*, which is less sensitive to outliers. One approach is to identify a clean subset of the data that is free of outliers and to calculate the scores of outliers relative to the clean subset (Simonoff, 1984). Given a subset x_{out} of outliers in x, another subset of x that excludes the outliers can be defined

$$x_{clean} = x \setminus x_{out}. \tag{24}$$

A modified Z score can be defined in terms of the clean estimates of the expectation value and standard deviation

$$Z_i = \frac{x_i - E[x_{clean}]}{\sigma_{x_{clean}}}. \quad (25)$$

In practise we calculate the expectation value with a piecewise regression model, which is fitted to historical data with robust regression. The standard deviation is calculated after excluding the outliers from the sample, according to the formula given above. Since the number of outliers in a sample is usually unknown the upper bound, r, in the GESD outlier identification step can be incremented iteratively until no more outliers are found. An interesting approach for the identification and ranking of potential outliers with linear models that does not require knowledge about the number of outliers is described by Hadi & Simonoff, 1993.

Another modified Z score can be defined in terms of the median absolute deviation (Iglewicz & Hoaglin, 1993)

$$Z_i^{MAD} = \frac{0.6745(x_i - x)}{median(\,x_i - x\,)}, \quad (26)$$

where x is the median of the sample. Median absolute deviation (MAD) scores with a magnitude above 3.5 are considered to be potential outliers. This is a robust score because the median, and median absolute deviation are robust statistics. In general, the MAD score is preferred when the distribution can be skew, which sometimes is the case here. However, the concepts of mean and standard deviation are more widely known and straightforward to integrate with standard piecewise regression models of the expectation value. Therefore, we use the first definition, Z_i, of the modified Z score in this work. This is also the score proposed by Seem, 2007.

4.3 Interpretation of Z scores

In order to understand qualitatively what a modified Z score of an outlier means in terms of probabilities it is instructive to consider the normal distribution. For a normally distributed random variable about 68% of values are within one standard deviation, σ, from the mean; about 95% are within two standard deviations; and about 99.7% are within three standard deviations. This means that most of the observed values of a normal distribution are within three standard deviations, which is a rule-of-thumb known as the *3-sigma rule*. More precisely, the probability that a normally distributed random variable lies outside the range $\mu \pm n\sigma$ is

$$P(\,x - \mu\, > n\sigma) = 1 - \mathrm{erf}\,\frac{n}{\sqrt{2}}\,, \quad (27)$$

where erf() is the error function (Olver et al., 2010). This relationship is illustrated in Figure 4.1. This relationship can be translated into an expected frequency or rate of

observations that are outside a certain number of standard deviations from the mean, see Table 4.1.

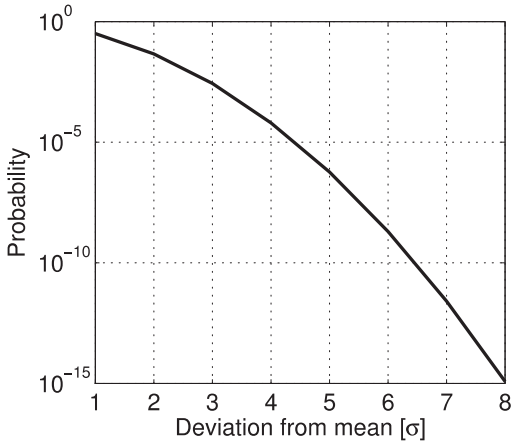


Figure 4.1. Probability that an observed value from a normal distribution deviates more than some number of standard deviations from the mean.

Table 4.1. Frequency of observations of a normally distributed variable that fall outside a certain number of standard deviations from the mean.

Deviation	Z score, Z_i	Average frequency
$\mu \pm 1\sigma$	≥ 1	1 in 3
$\mu \pm 2\sigma$	≥ 2	1 in 22
$\mu \pm 3\sigma$	≥ 3	1 in 370
$\mu \pm 4\sigma$	≥ 4	1 in 15 787
$\mu \pm 5\sigma$	≥ 5	1 in 1 744 278
$\mu \pm n\sigma$	$\geq$ n	1 in $\frac{1}{1-\mathrm{erf}\,\frac{n}{\sqrt{2}}}$

In reality the hourly data that we are modelling here does not have residuals that are normally distributed. For example, the normal distribution overestimates the tail probabilities because physical limitations of the substation prevent arbitrarily high and low loads. There are also systematic errors, which are consequences of the simplicity of the model and the incomplete knowledge about the processes that affect the overall energy use. Skew distributions of the residuals are common, but some substations have nearly symmetric distributions. Anyway, the proposed scores are useful for the identification of outstanding outliers using a ranking procedure. This point is illustrated by the empirical test results that are presented below. An example of outliers identified in data from an apartment building is illustrated in Figure 4.2.

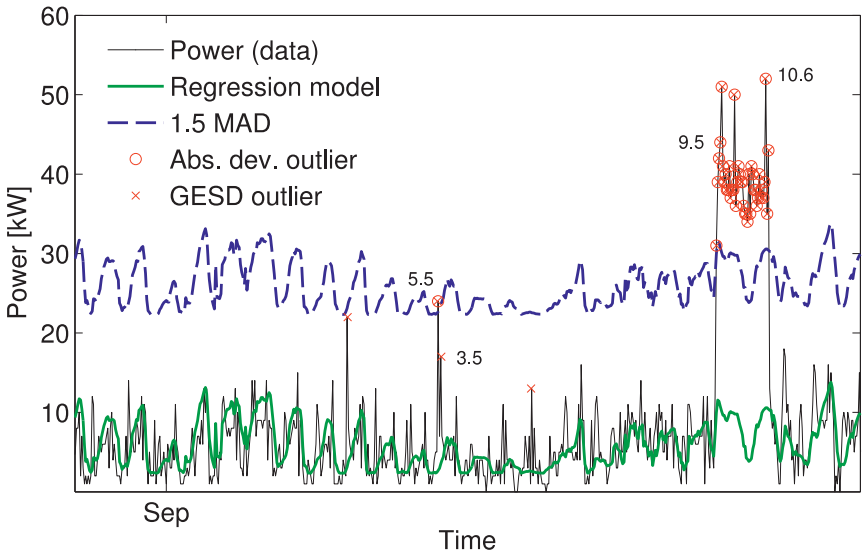


Figure 4.2. Outliers identified with the GESD method and Z score ranking in historical data from an apartment building. These abnormal values were not discovered during ordinary operation and their cause is unknown. The piecewise regression model (bold solid line) and maximum absolute deviation (MAD) with a 50% added tolerance (bold dashed line) is displayed in addition to the data (thin solid line). Data identified as potential outliers with the GESD method (crosses) and data that falls outside the 1.5 MAD limit (circles) are highlighted. The numbers next to the highlighted outliers denote modified Z scores $(3.5, 5.5, 9.5, 10.6)$.

4.4 Complementary ranking methods

The modified Z score that is introduced above is a useful but not sufficient indicator of anomalies. In particular, if the data record of a substation includes faults from the beginning, or during the reference period used to calculate the regression model it is possible that the Z scores are low even though data is faulty. In such cases the faulty behaviour is the norm and it cannot be identified as abnormal.

The absolute deviation, $x_i - E[x]$, is a useful complement to the Z score because it is an absolute quantity that is independent of the expected variance of the variable. A corresponding relative measure that does not give preference to high-load customers can be defined by normalizing the absolute deviation with the contracted power, or median power for each substation. Therefore, in addition to Z-score ranking, substations can be ranked according to the maximum absolute deviation (MAD). Given a model for calculation of energy cost this approach can also be extended to include the economic risk that is associated with absolute and relative deviations, but we have not considered that possibility in this work. In addition to the magnitude of a Z score, the duration and frequency of outliers (Pakanen et al., 1996, Section 3.7) or the sum of Z scores of outliers can be considered. In principle, a

learning-to-rank approach (Murphy, 2012, Section 9.7) could be developed so that the ranking mechanism is improved when faults are identified and confirmed.

4.5 Test results

We calculate the maximum modified Z scores for the population of 996 substations in the test set. The calculation is based on one year of data and we consider the power, the flow and the primary supply temperature separately in the three subsections below. We display the annual power profiles, flow profiles and supply temperature profiles for substations that have outliers with exceptionally high Z scores. Note that (given a regression model, which needs to be fitted to historical data), the outlier detection test and the calculation of a modified Z score can be performed online when an hourly value is received from an energy meter. Therefore, outliers with high Z scores can be detected immediately.

4.5.1 Detection of abnormal power

We calculate the maximum magnitude of modified Z scores of the power data for all 996 substations in the data set, using regression models with eight piecewise linear segments that are automatically fitted to the data; see Figure 4.3.

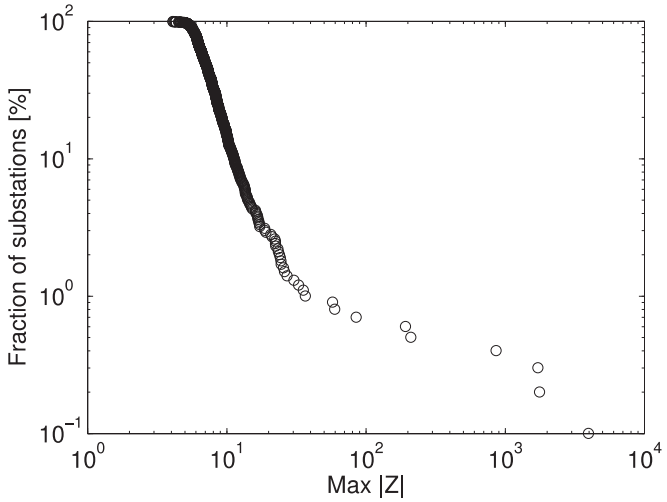


Figure 4.3. Maximum magnitude of modified Z scores for all substations in the test set. There are six substations that have at least one modified Z score above 100.

The modified Z scores are calculated using a robust estimate of the standard deviation, which means that outliers are excluded and that the standard deviation is calculated from the remaining residuals of the regression model. Next we present the annual power profiles of the 18 substations that have the highest magnitudes of Z scores; see Figures 4.4 – 4.6. Faults have been detected and addressed in most of these cases.

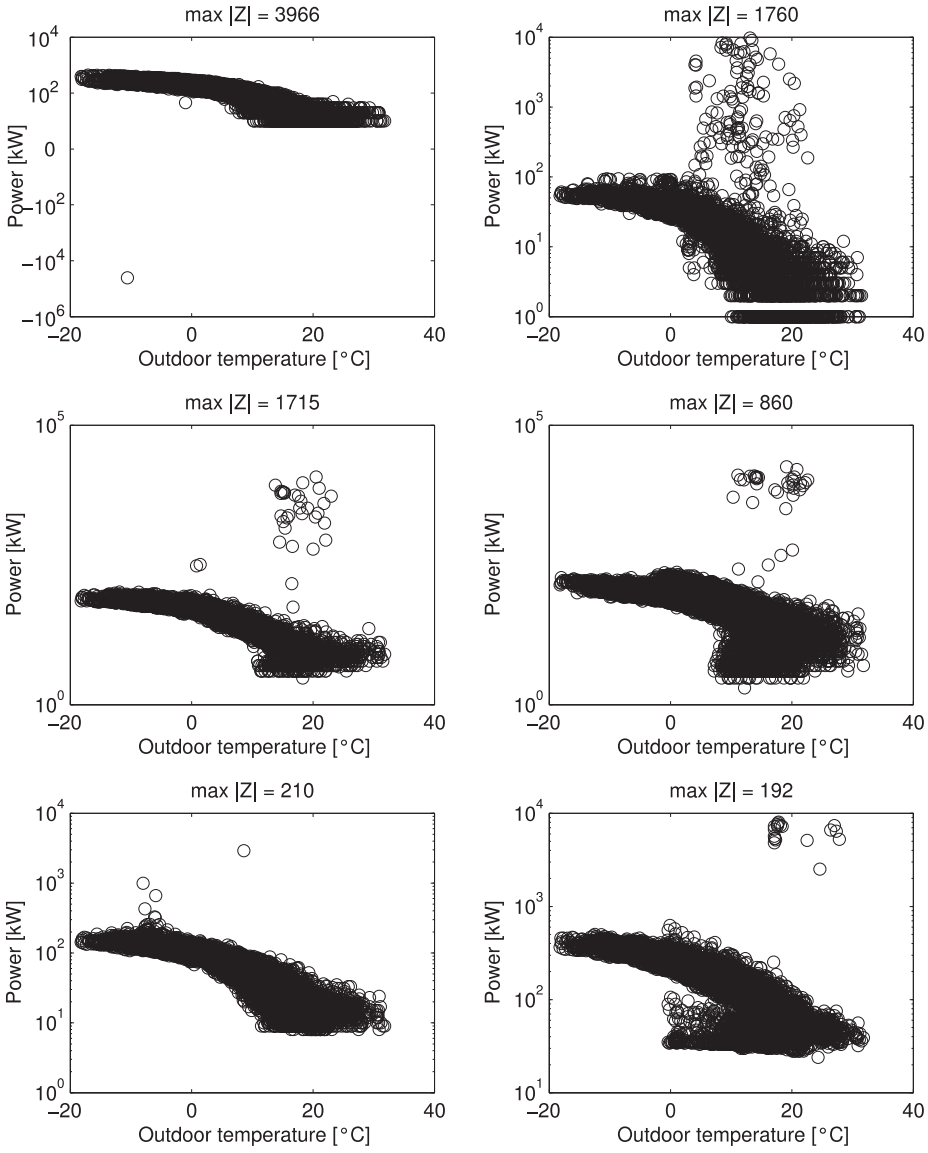


Figure 4.4. Power profiles of the six substations with the highest Z scores.

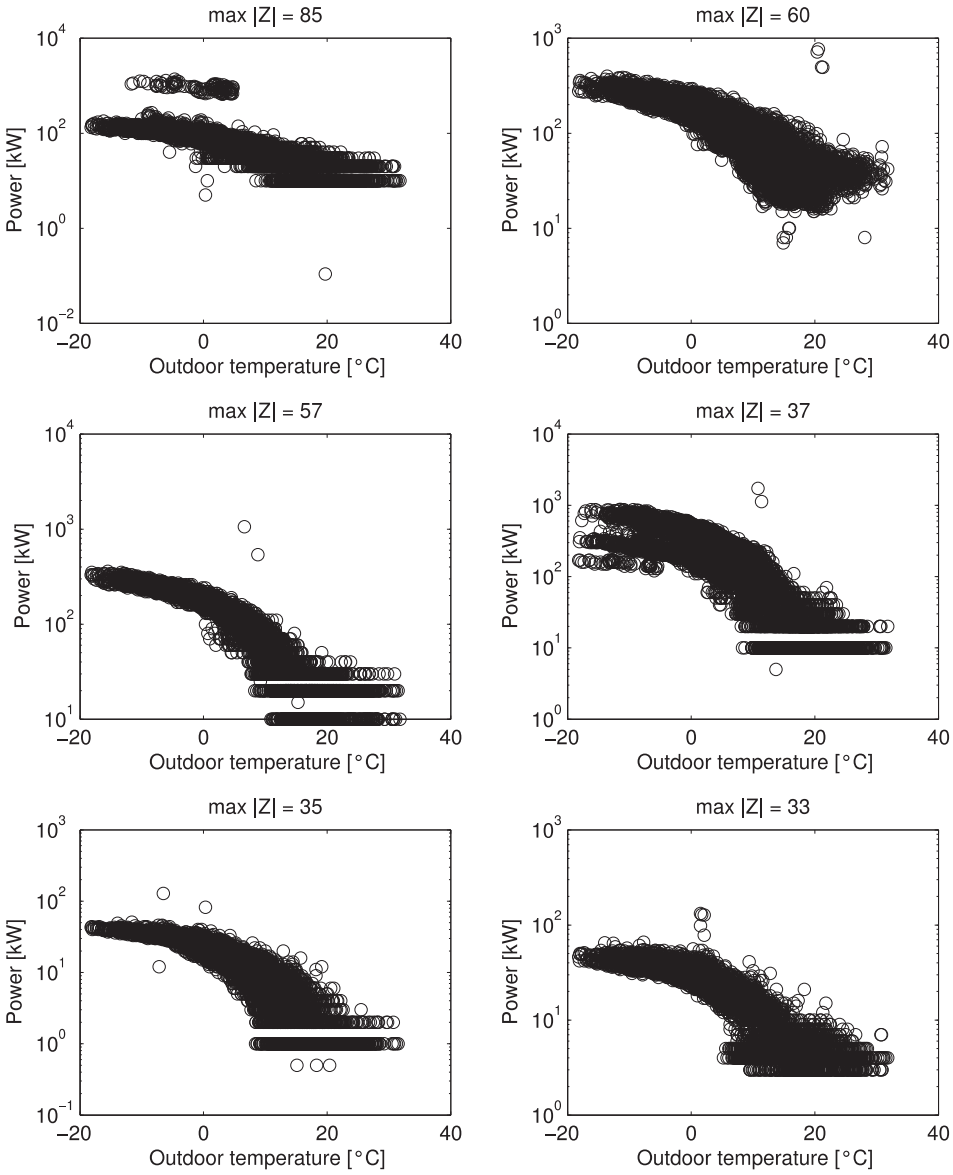


Figure 4.5. Power profiles of the top 7–12 substations with high Z scores.

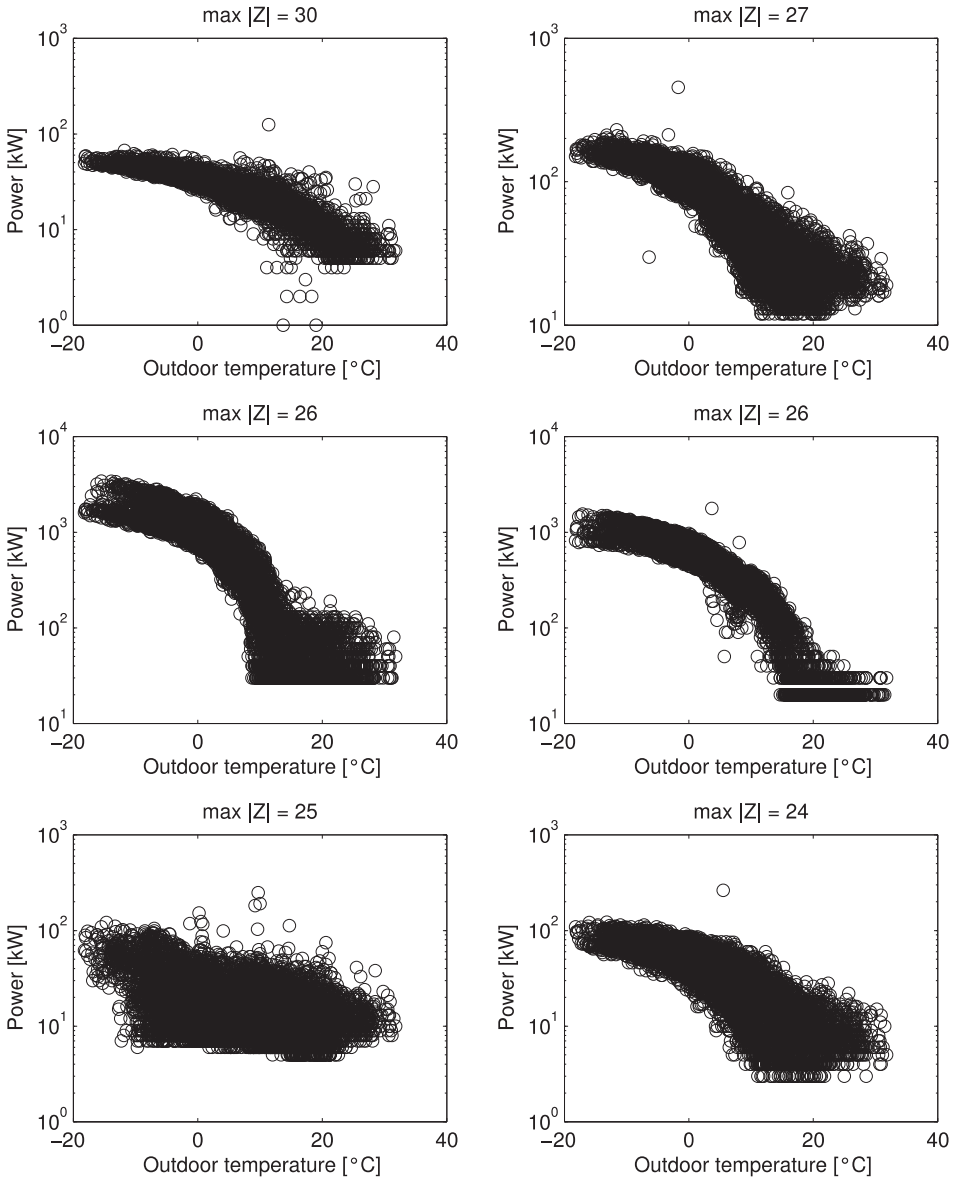


Figure 4.6. Power profiles of the top 13–18 substations with high Z scores.

4.5.2 Detection of abnormal flow

Outliers in the flow data are detected much like outliers in the power data. Piecewise linear models with eight segments are automatically fitted to the flow profiles of the substations and the GESD outlier test is applied to the model residuals. The maximum magnitudes of the Z scores for the 996 substations in the population are displayed in Figure 4.7. Next we present the annual flow profiles of the 18 substations that have the highest magnitudes of Z scores; see Figures 4.8 – 4.10.

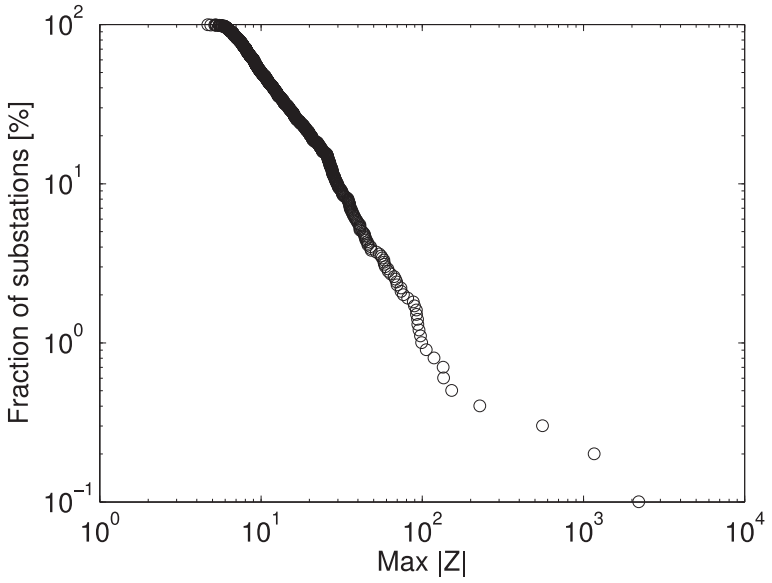


Figure 4.7. Maximum magnitude of modified Z scores for the flow data of all substations in the test set. There are nine substations that have at least one modified Z score above 100.

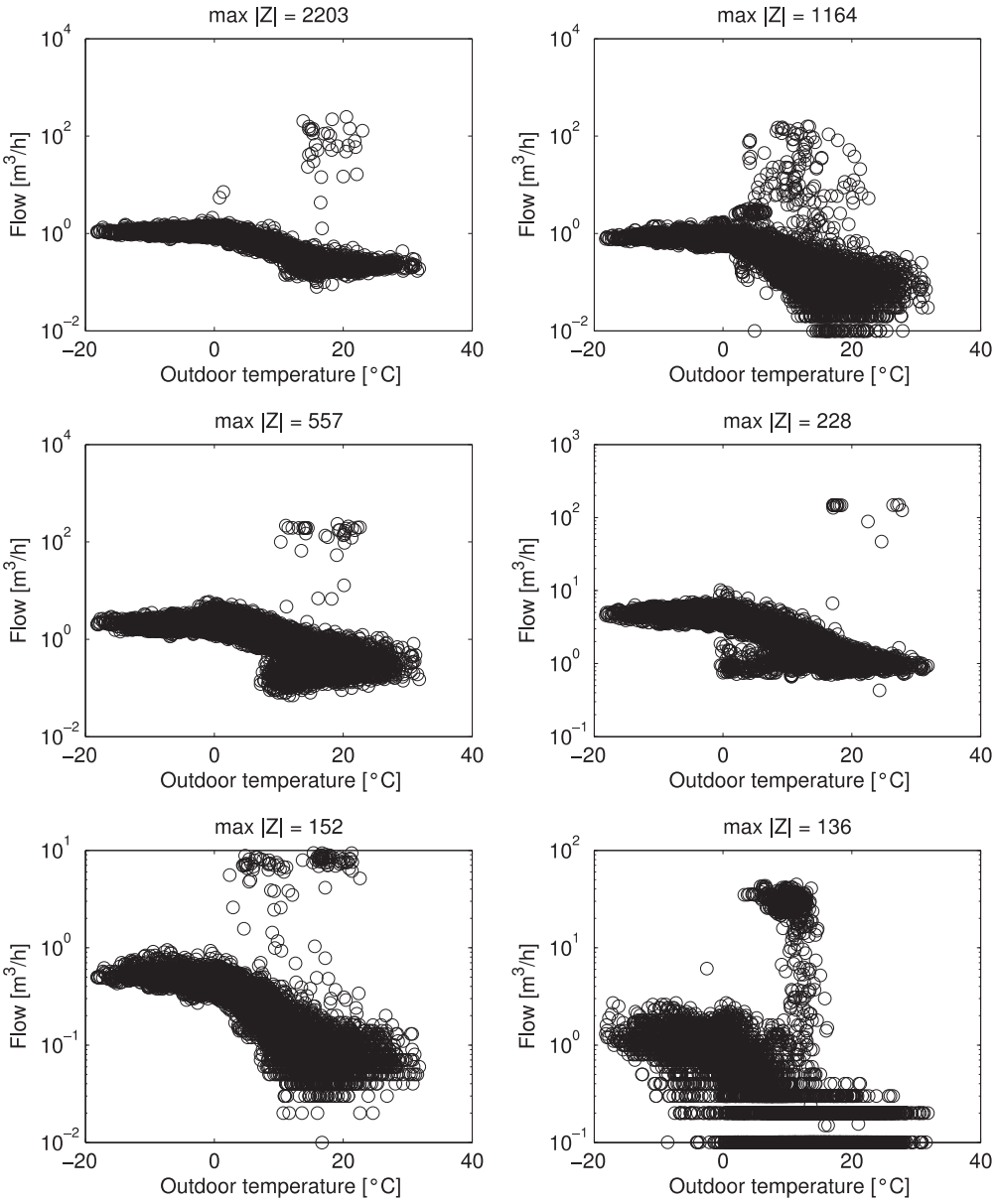


Figure 4.8. Flow profiles of the six substations with the highest Z scores.

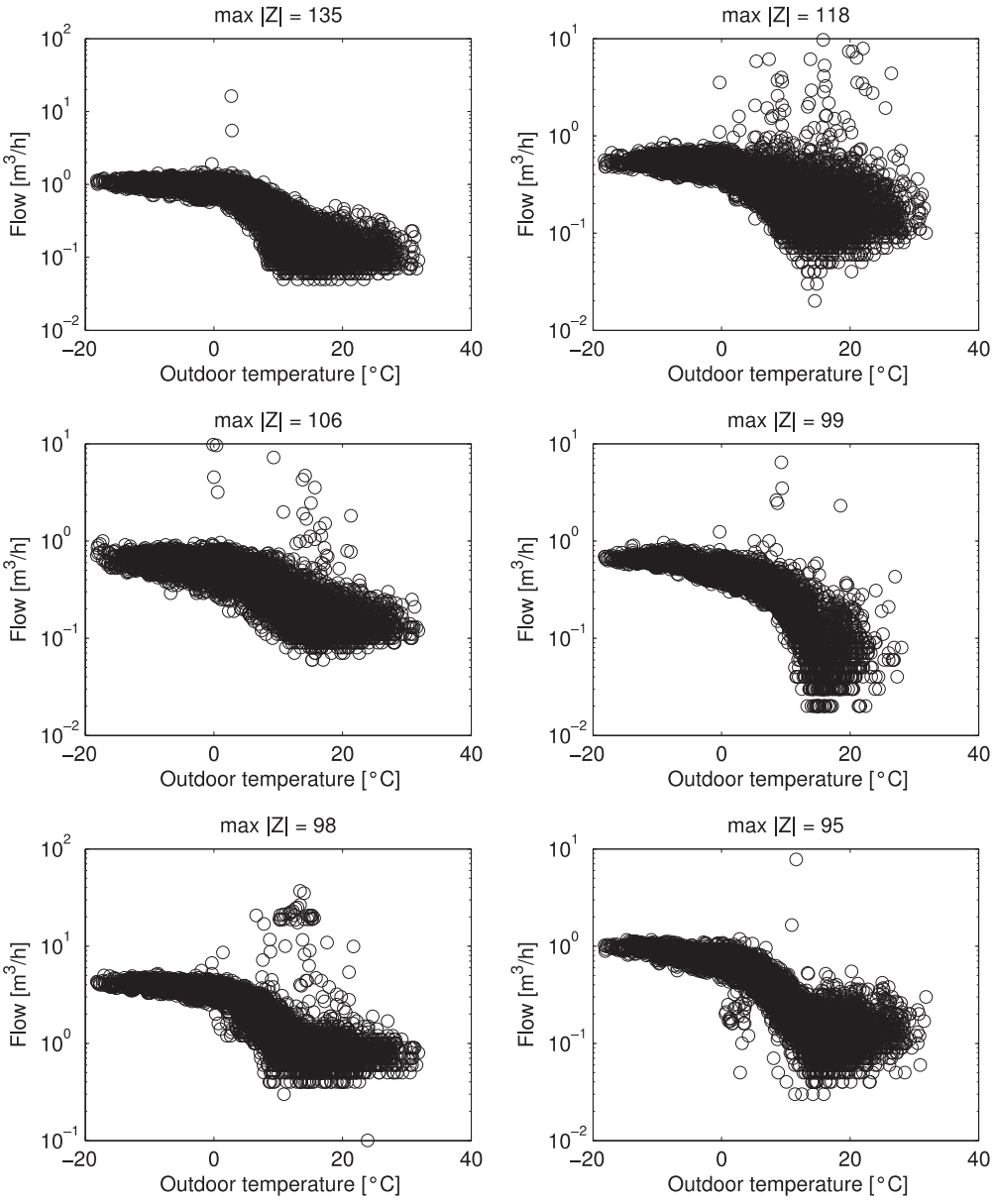


Figure 4.9. Flow profiles of the top 7–12 substations with high Z scores.

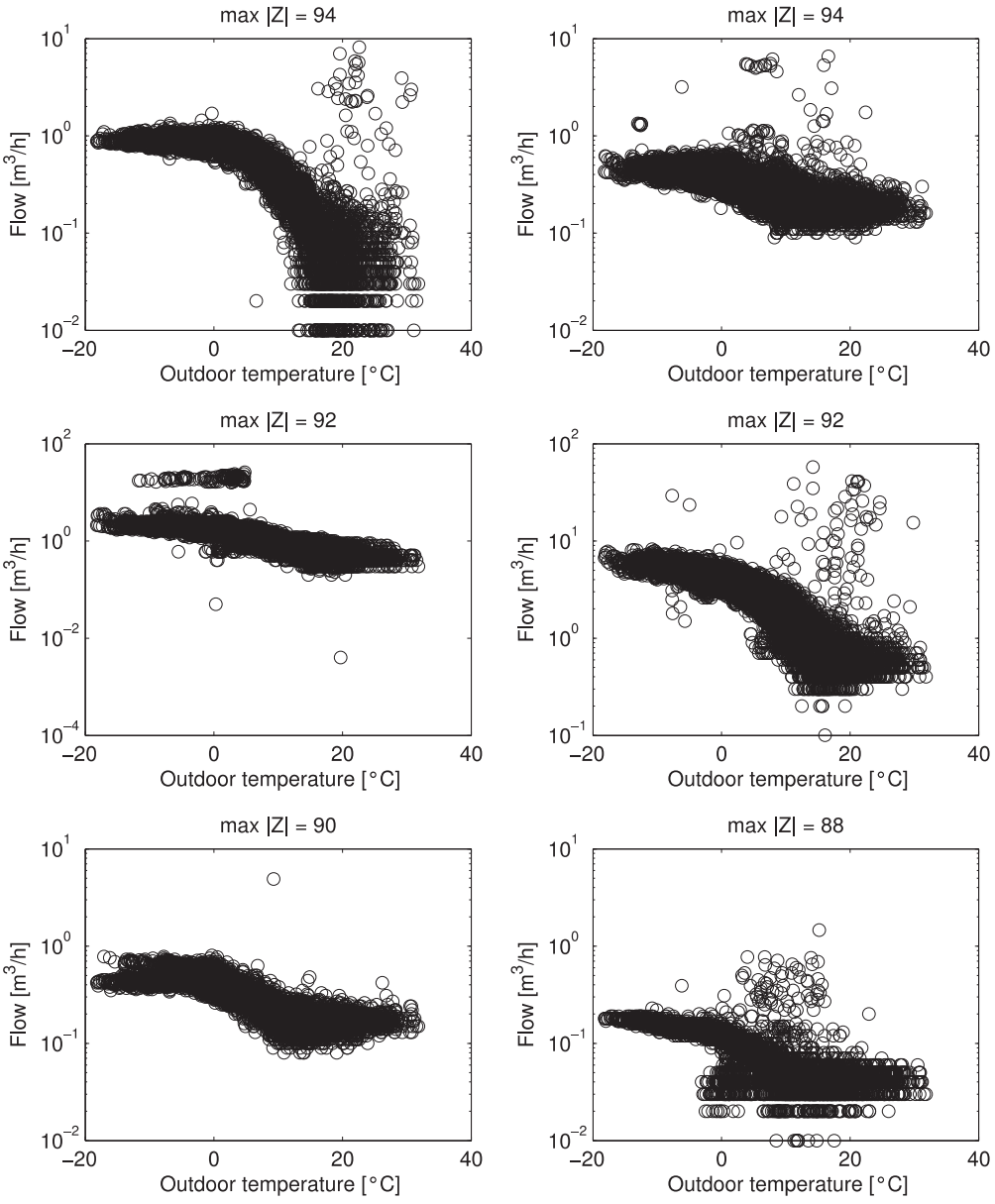


Figure 4.10. Flow profiles of the top 13–18 substations with high Z scores.

4.5.3 Detection of abnormal supply temperatures

The primary supply temperature is analysed in a different way compared to the power and flow data. For each substation, correlation analysis is used to identify another substation with a maximally correlated de-trended supply temperature. It is not necessary to perform bimodality and cluster analysis in this case. Piecewise linear models are automatically fitted to the primary supply temperature difference, ΔT_{ps}, of each pair of substations versus the geometric mean of the primary flow of the two substations. The GESD outlier test is applied to the regression model residuals, just like we did for the power and flow. The maximum magnitudes of the Z scores for the 996 substations in the population are displayed in Figure 4.11. Next we present the annual supply temperature profiles of the 18 substations that have the highest magnitudes of Z scores; see Figures 4.12 – 4.14.

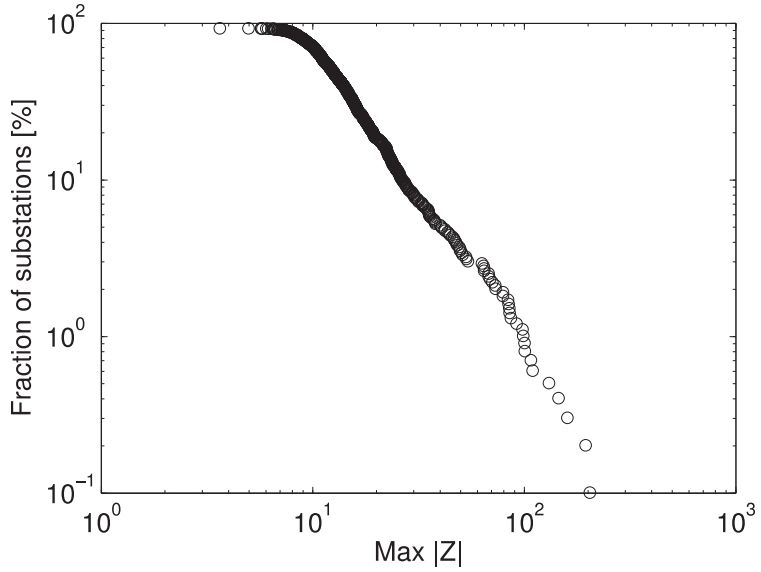


Figure 4.11. Maximum magnitude of modified Z scores for the supply temperature of all substations in the test set. There are nine substations that have at least one modified Z score above 100.

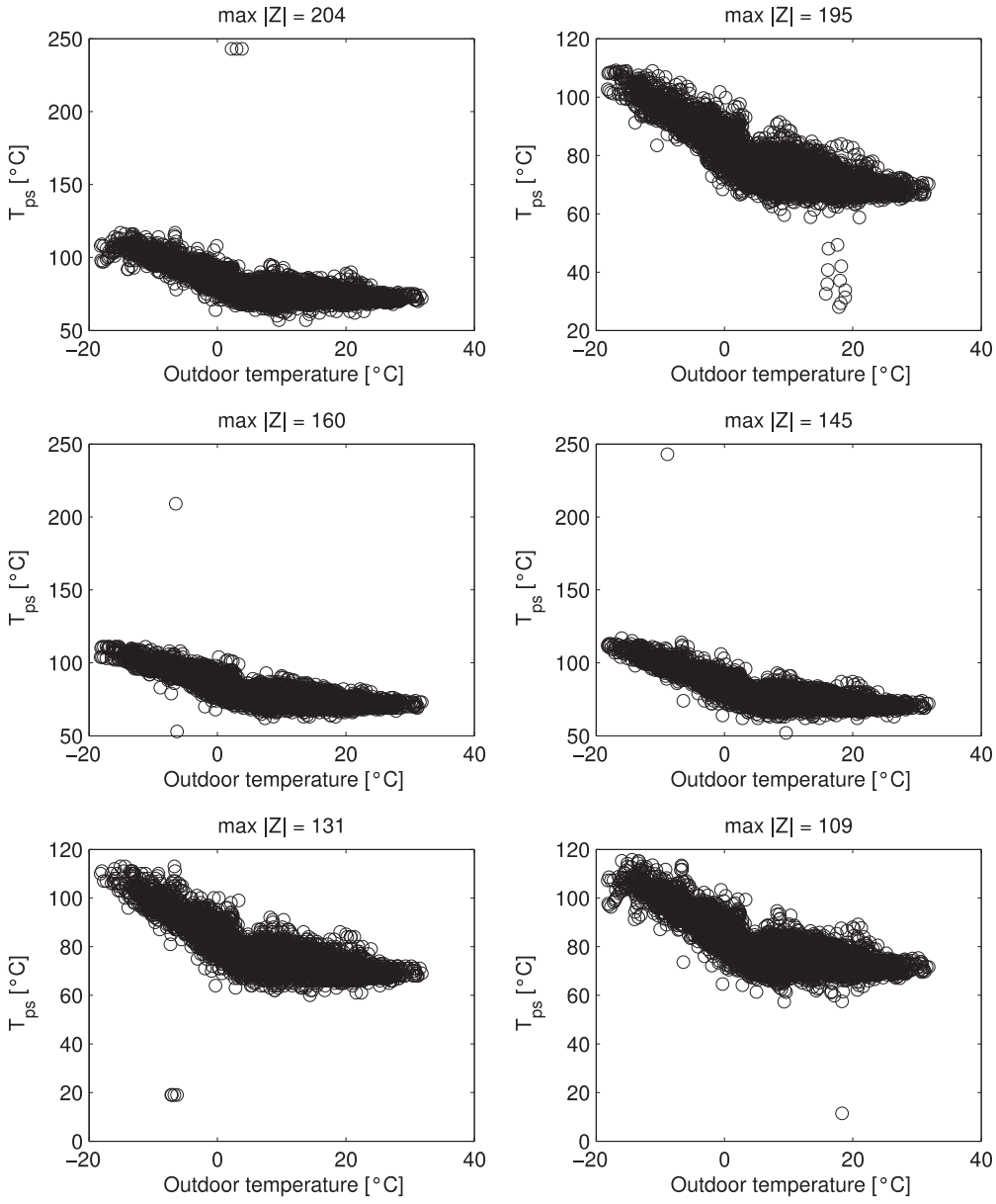


Figure 4.12. Supply temperature profiles of the six substations with the highest Z scores.

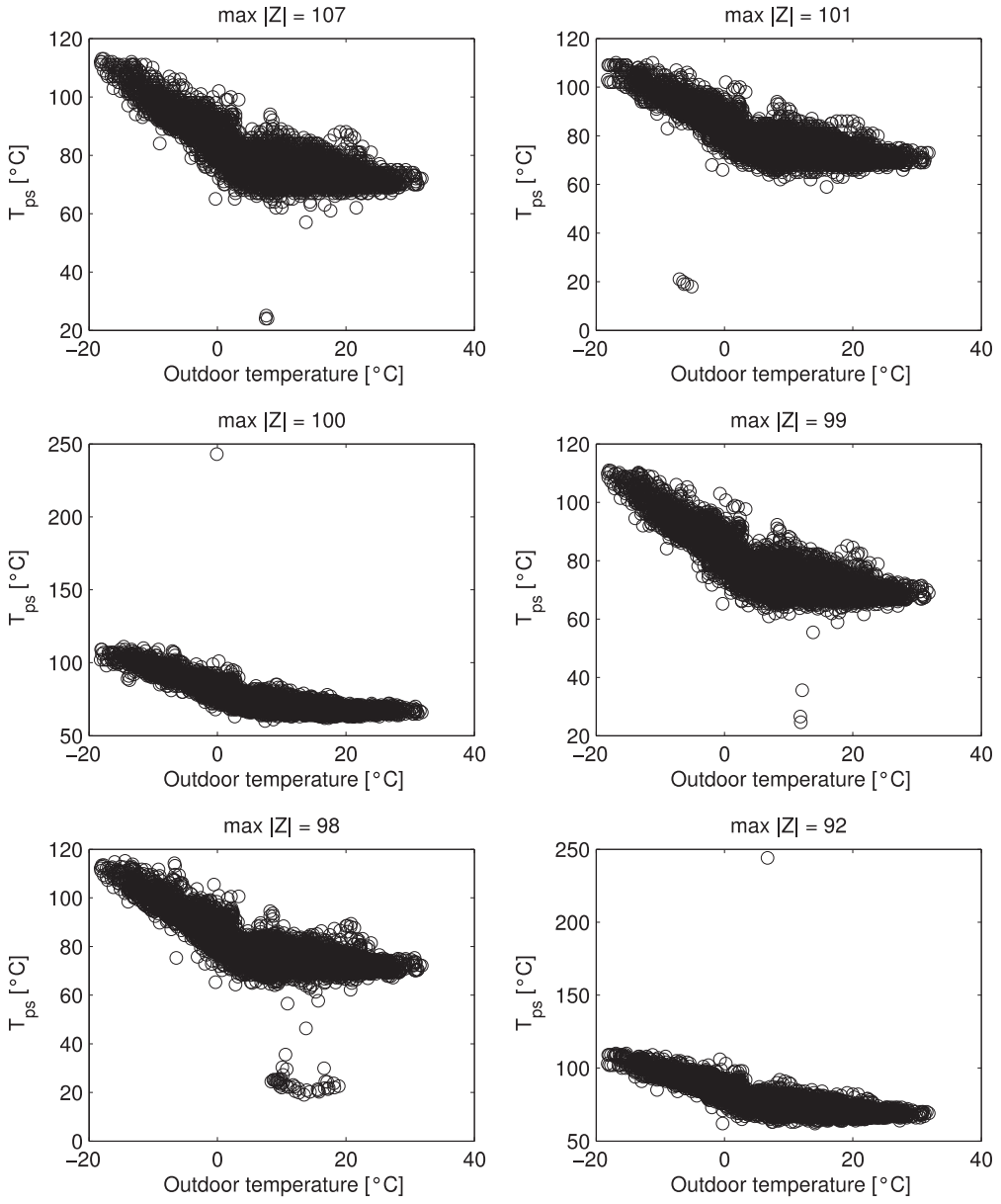


Figure 4.13. Supply temperature profiles of the top 7–12 substations with high Z scores.

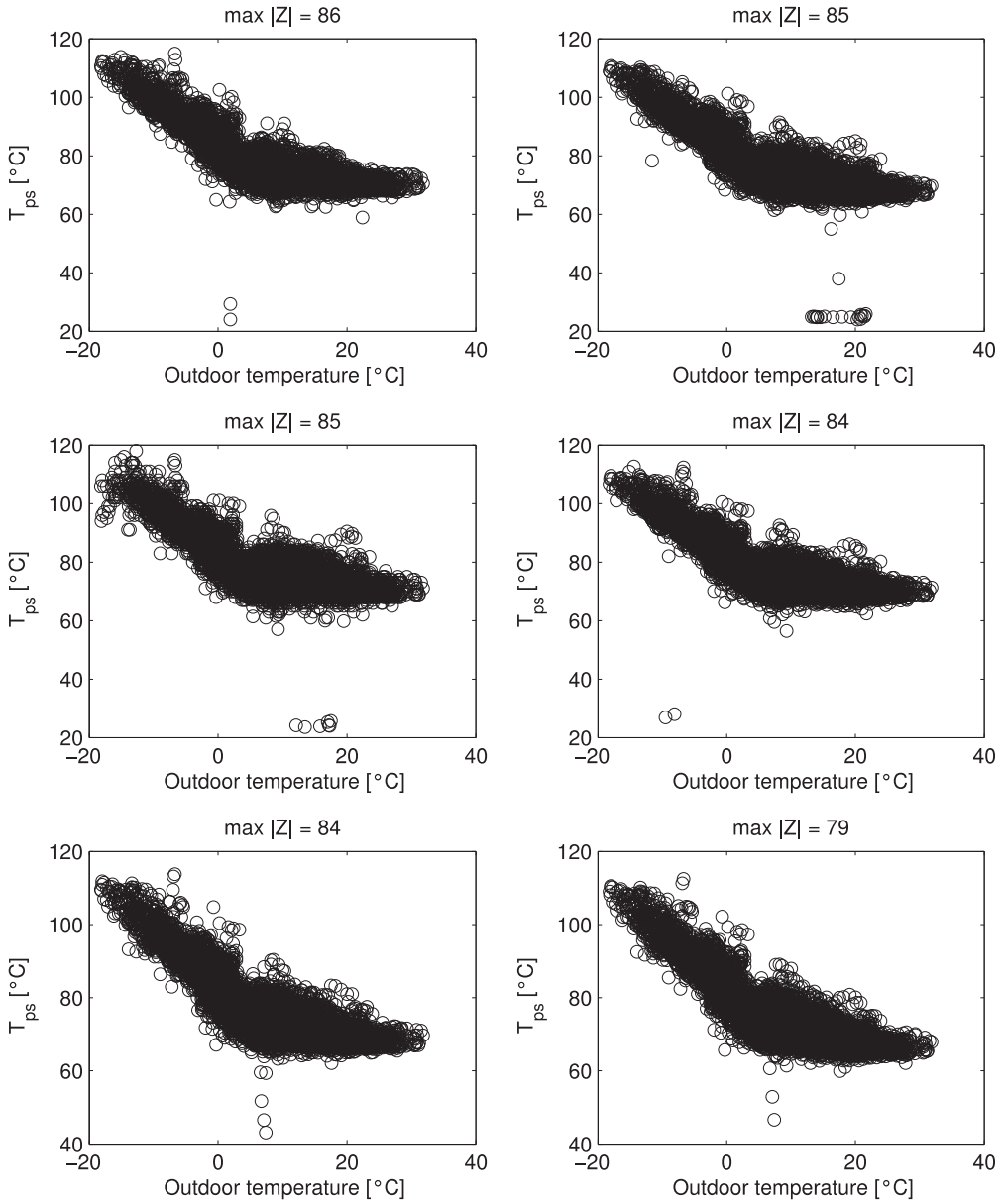


Figure 4.14. Supply temperature profiles of the top 13–18 substations with high Z scores.

4.5.4 Comparison of results

We compare the top-lists of the maximum Z scores of different substations, which are illustrated in Figures 4.3, 4.7 and 4.11, in order to see whether faulty substations are identified, and whether some faulty substations appear on more than one top-list. The result is illustrated in Table 4.2. Seven out of the top-10 substations with abnormal outliers in the power data have been confirmed as faulty. The status of the remaining three is unknown. Included in the table is also the position of these ten substations in the top-lists of Z scores for outliers in the flow and primary-supply temperature data, respectively. Most of the ten substations are highly ranked on at least two top-lists, and one substation is among the top-11 on all three lists. These results indicate that substations that appear high up on the top-lists of outliers with high Z scores should be investigated because they are likely to be faulty.

Table 4.2. Summary of the top-10 substations displayed in Figure 4.3, which have the highest maximum Z scores of outliers in the power data. The indices / positions of these ten substations in the top-list displayed in Figure 4.7 is included in the second column. The third column displays the indices of these ten substations in the top-list illustrated in Figure 4.11. The fourth column indicates whether a fault is identified.

Top-list position			
Power, max |Z| (cf. Figure 4.3)	Flow, max |Z| (cf. Figure 4.7)	T_{ps}, max |Z| (cf. Figure 4.11)	Fault confirmed
Top-1	Top-323	Top-329	Yes, energy meter replaced.
Top-2	Top-2	Top-113	Yes, flow meter and temperature sensors replaced.
Top-3	Top-1	Top-262	Yes, instrumentation rebuilt.
Top-4	Top-3	Top-380	Yes, energy meter and flow meter replaced.
Top-5	Top-191	Top-3	No, cause unknown.
Top-6	Top-4	Top-638	Yes, instrumentation rebuilt.
Top-7	Top-15	Top-451	Yes, energy meter, flow meter and temperature sensors replaced.
Top-8	Top-272	Top-554	Yes, communication device replaced.
Top-9	Top-11	Top-7	No, cause unknown.
Top-10	Top-103	Top-67	No, cause unknown.

4.6 Implementation

Ranking of outliers can be implemented online in an energy-meter data management system, for example in the form of a sorted list of substations that is automatically updated. This list should have multiple columns with associated sorting functions, including a column for the modified Z score, the absolute deviation and the normalized absolute deviation. Plots similar to Figures 4.3 and 4.11 illustrate the

distribution of modified Z scores for the whole population of substations, which is helpful to identify exceptional outliers. Such plots can be created also for the absolute deviation and the normalized absolute deviation.

A plot of time sequence data and outliers similar to that in Figure 4.2 is helpful for manual diagnosis. A plot of that type could for example be displayed when a substation is selected in the sorted list. The interface used to calculate and display regression models should be easily accessible so that regression models from different reference periods can be selected in order to see the effect on the outliers and Z scores. In addition to outliers in the power, outliers in the flow and ΔT_{ps} can be used for diagnosis purposes to analyse the cause of potential faults.

A method for tagging of outliers is needed so that potentially faulty substations can be monitored. Tags can also be used to notify operators when a regression model is calculated for a time period that contains outliers or faulty data.

5 DRIFT DETECTION

There are several components of an energy meter that can change behaviour over time, in particular the flow sensor, temperature sensors and electronics (amplifiers and voltage references). Therefore, faults can emerge gradually under long-term operation, leading to an increasing bias in the energy calculated by the energy meter. We refer to this behaviour as *drift*. This type of fault is difficult to detect with the basic limit checking and outlier analysis methods that are described above because there is no sudden, abnormal change of the variables. Because faults like these can be integrated over long periods of time the associated cost can be high. Therefore, a method for detection of drift is described here.

5.1 Illustration with regression models

The following example illustrates the problem that we are addressing. If there is a significant change in the mean power from one period of time to another the resulting offset can be illustrated by comparing two regression models, which are fitted separately to the data from each period of time, see Figure 5.1.

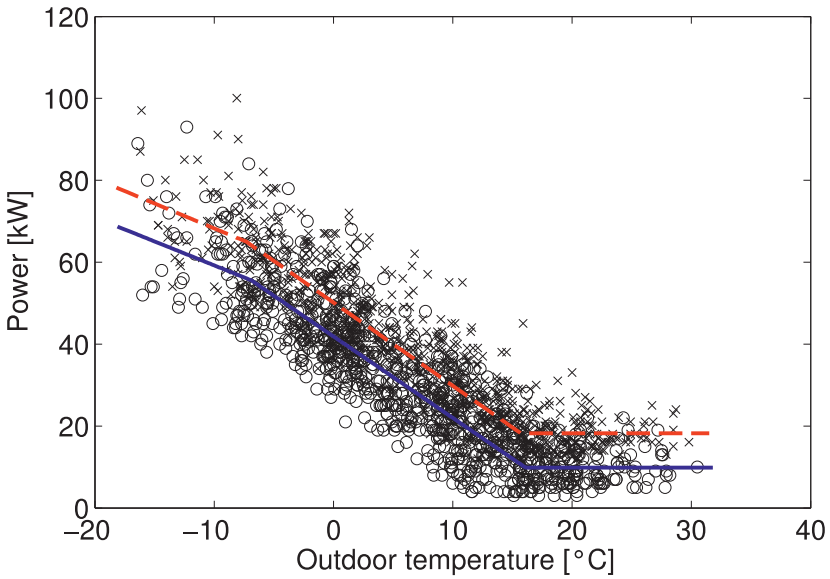


Figure 5.1. Regression models fitted to data from a substation at two different time periods. The power is artificially biased by one standard deviation (8 kW) during one of the time periods (dashed line, crosses), which results in an offset compared to the unbiased case (solid line, circles).

This method is useful for visual inspection, and as a pedagogic tool for the communication with customers when faults are detected and payments have to be adjusted. Note that regression models can change over time also in the absence of

faults, for example when buildings or building subsystems such as the ventilation are modified or upgraded. Next we introduce a method for automated detection of drift.

5.2 Drift detection with cumulative sums

From a mathematical point of view faults of this type are associated with a gradual change of the probability density function of the residuals of the process model, in particular the mean. A *cumulative sum* of residuals is a simple and effective statistic for early detection of small changes in the mean, which can be implemented in the form of cumulative sum control charts (Woodall & Adams, 1993; NIST, 2012), so-called CUSUM charts. A cumulative sum, S, can be defined iteratively in this way

$$S_{i+1} = S_i + \frac{1}{\alpha}(x_i - E\ x_i\ ), \quad S_0 = 0, \tag{28}$$

where x_i are the values of a variable with expectation value $E\ x_i\ $, for example the thermal power, and α is an optional normalization parameter. When new values become available the difference between each value and the expectation value is summed in a cumulative manner. If the expectation value is correct the deviations from the expectation value will average. On the other hand, if the values are sampled from a distribution with a biased mean value the magnitude of the cumulative sum will progressively increase.

If the change of the mean can be either positive or negative, which is the case considered here, the cumulative sum statistic is better divided in two parts

$$S_{i+1}^{+} = \max\ \ 0,\ S_i^{+} + \frac{1}{\alpha}(x_i - E\ x_i\ - k)\ \ ,\quad S_0^{+} = 0, \tag{29}$$

$$S_{i+1}^{-} = \max\ \ 0,\ S_i^{-} - \frac{1}{\alpha}(x_i - E\ x_i\ + k)\ \ ,\quad S_0^{-} = 0. \tag{30}$$

Here k is a *reference level* for the magnitude of the shift in the mean that we wish to detect. If the probability density function of x is known the probability that the magnitude of a cumulative sum exceeds some value can be calculated and a test statistic can be defined (NIST, 2012, Section 6.3.2.3). In that case a decision limit, h, is defined and whenever S_i^{+}or S_i^{-} is equal to or greater than h the mean of the variable is considered to be shifted. A typical choice of parameters is to set $\alpha = \sigma_x$, which means that the statistic is *standardized*, and k is set to a fraction of the standard deviation, σ_x. The test statistic can perform poorly when the size of the mean shift is significantly different from the assumed reference level. One approach to address that problem is to assign a probability density to the reference level (Ryu et al., 2010).

When modelling a large population of district energy substations automatically with piecewise regression and hourly energy metering data the probability density functions of the residuals varies between substations, are not normally distributed, and are often skew. Therefore, it is difficult to define definite decision limits for change detection. However, the effect of a gradually increasing change of the mean

eventually becomes significant and can be detected. An example is illustrated in Figure 5.2. In this figure the power is artificially shifted with time in a linear fashion so that the shift is zero at the beginning of the displayed time period and one standard deviation, $\sigma_P = 8$ kW, at the end of the one-year time period. Other parameters are $\alpha = \sigma_P$ and $k = 0.25\sigma_P$. The power profile of this substation is illustrated in Figure 5.1 and the non-shifted regression model that is displayed in that figure is used to calculate the expectation value of the power, $E\ P$.

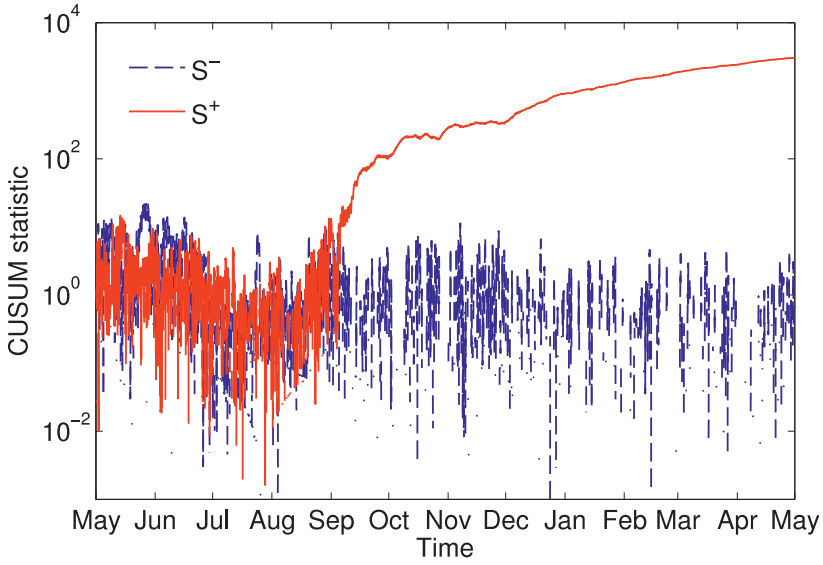


Figure 5.2. Cumulative sum statistic of the power, P, versus time in the presence of a gradually increasing bias in the power.

The statistic S_i^+ starts to deviate in September, when the shift of the power is about 3 kW. The two regression models that are illustrated in Figure 5.1 represent the models that result before and after the introduction of the one-σ_P change of the power. The offset between the two regression models is about one standard deviation, which is a change of order 10%. In contrast, the CUSUM statistic S_i^+ changes by more than one order of magnitude during the same time period. Next we describe how this statistic can be used to identify abnormal substations in a population.

5.3 Ranking with cumulative sums

The cumulative sum of different substations can be qualitatively compared and ranked if the normalization constant, α, is chosen so that the magnitudes of the sums are comparable. This can be achieved if a standardized variable with normalization constant $\alpha = \sigma$ is used. The reference level, k, should not be set too high or too low because then changes can remain undetected, or the cumulative sums can be high by chance, respectively. Since we are interested predominantly in small and gradually increasing changes of the mean, the reference level should be comparable to the

standard deviation, $k \sim \sigma$. The cumulative sums S_i^+ and S_i^- can be calculated for each substation, starting at the end of some reference time period with associated known regression models of the data. The substations can be scored with a maximum cumulative sum statistic

$$CS = \max\ S_i^+,\ S_i^- \ , \tag{31}$$

so that substations with outstanding cumulative sums can be identified with a ranking procedure. Depending on the magnitude and historical trend of the cumulative sum statistic the top-ranked substations can be further investigated or tagged for continued monitoring. Figure 5.3 illustrates an example of how the CS statistic can be affected by an increasing bias of the power, which is simulated here with an additive linear function of time. The bias term increases from zero at the beginning of the time period to one (or one half) standard deviation at the end of the time period, as indicated by the legends in the figure.

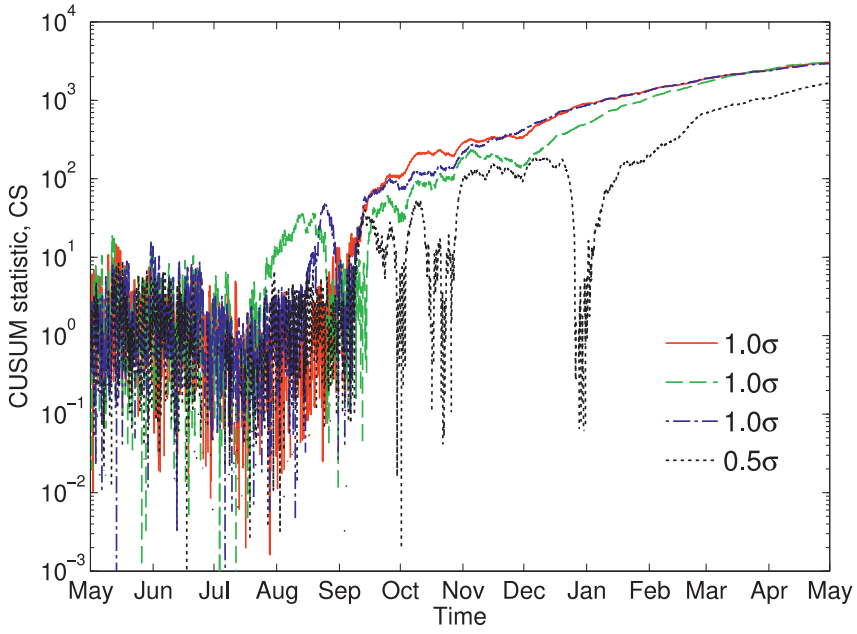


Figure 5.3. Cumulative sum statistic for the power of four different substations. In three cases the power is shifted by a bias that increases linearly with time from zero to one standard deviation, resulting in comparable CS statistics at the end of the one-year time period. In the fourth case the power drifts by one half standard deviation, resulting in a lower CS value.

The CUSUM parameters are $\alpha = \sigma_P$ and $k = 0.25\sigma_P$. The drift of the power results in a long-term trend towards higher values of CS. In all four cases the CS statistic increases with time after a threshold value of the drift is reached, and it exceeds 10^3 at the end of the one-year time period.

In general, a cumulative sum of a random variable can attain high values by chance. For example, the cumulative sum of a normally distributed random variable with unit variance will attain values higher than 10^3 with high probability after summing a few million samples. The purpose of the reference level, k, in the CUSUM statistic is to supress such random fluctuations. If the reference level, k, is low compared to the standard deviation the CS statistic can become high also in the absence of a bias term because the residual of the regression model is a random variable. In order to illustrate the effect of a varying reference level we calculate the hourly CS statistics over a period of one year for 853 substations in the dataset with BC below 0.6 (see Section 3.2 for further information about this selection criteria). Figure 5.4 displays the relative number of substations in the population that have at least one CS value as high as that indicated by the horizontal axis. For example, with $k = 0.25\sigma_P$ all substations (100%) have at least one CS value of 10, while only 80% of the substations have CS values of 10^2, and less than 10% of the substations have CS values of 10^3. These numbers are lower for higher values of k. For example, with $k = \sigma_P$ less than 20% of the substations have CS values of order 10^2.

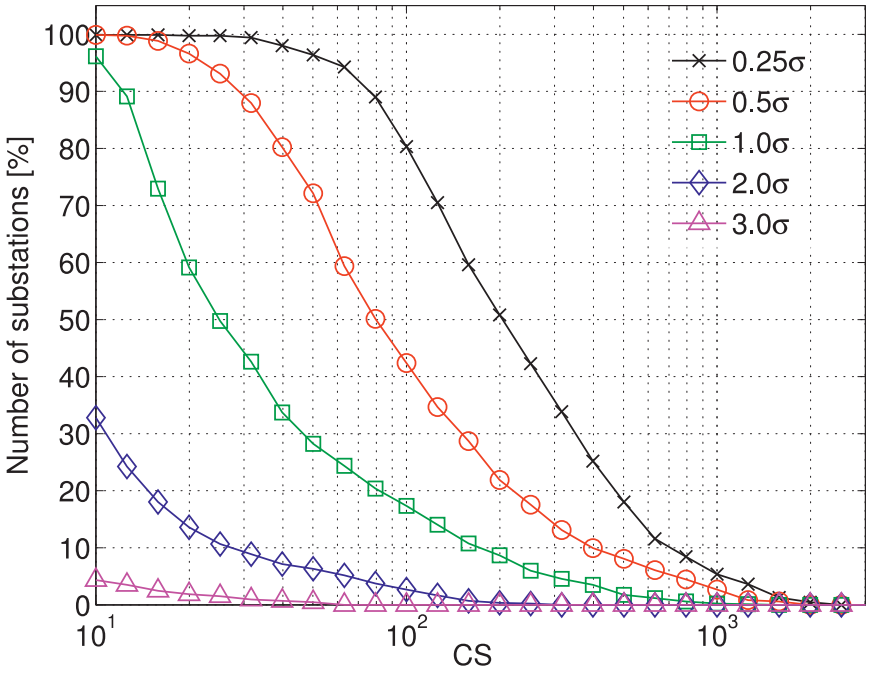


Figure 5.4. The relative number of substations in a population versus the CS values attained for five different choices of the reference level, k. In this calculation no bias term is added to the power.

In this calculation no bias term is added to the power and the data is pre-processed with the GESD method (see Chapter 4) in order to reduce the effect of faults that appear as outliers in the dataset. The high CS values in Figure 5.4 result from the real-

world residuals of the 853 regression models. This figure shows that CS values of 10^2 are common for reference levels below one standard deviation, while less than 10% of the substations have CS values approaching 10^3 during the same time period. Few substations have CS values exceeding 10^3, which means that the high values attained in the example illustrated in Figure 5.3 can be identified with a ranking procedure. We find that the high CS values in Figure 5.4 typically are associated with short-term peaks that are characteristically different from the monotonous long-term trends that are illustrated in Figure 5.3. Therefore, visual inspection of the trends of the relatively few substations that have exceptionally high CS values can be carried out to identify and tag potentially faulty substations with long-term drift for continued monitoring or field inspection.

In principle this method can be generalized to other variables, provided that there is a (regression) model that can be used to calculate the expectation value of the variable. For example, the CS statistic of the flow, m, and supply temperature difference, ΔT_{ps}, can be calculated using piecewise regression models (Chapter X).

5.4 Implementation

The maximum cumulative sum statistic CS is straightforward to integrate with the online interface for ranking of outliers (Section 4.6) in the form of an extra sortable column. An implementation of the CUSUM statistic with name `cusum` is provided in the appendix. A plot of time sequence data similar to that in Figure 5.3 is helpful for manual diagnosis. A plot of that type could for example be displayed when a substation is selected in the sorted list. The reference level should be optional, so that it can be adjusted for substations with exceptionally high variability in the residuals. The interface used to calculate and display regression models should be accessible so that regression models from different reference periods can be selected in order to see the effect on the CS, S_i^+ and S_i^- statistics.

6 DETECTION OF ABNORMAL QUANTIZATION

Oversized or faulty flow meters, malfunctioning flow valves, faults in the cabling or electronics, and rounding errors can lead to abnormal *quantization* of energy meter data, which means that the values of a variable are limited to a few discrete levels and have poor precision. For example, an oversized or incorrectly configured flow meter can result in insufficient precision of the flow measurements, and an old valve that is stuck in one position can result in constant flow. An example of a substation with poor precision in the hourly flow and power data is illustrated in Figure 6.1. These types of faults are difficult to detect with the outlier and drift detection methods that are described above because there is not necessarily an associated increasing bias or sudden, abnormal change of the variable. Because faults of this type can be integrated over long periods of time the associated cost can be high. Therefore, we describe a method for detection of abnormal quantization of variables here.

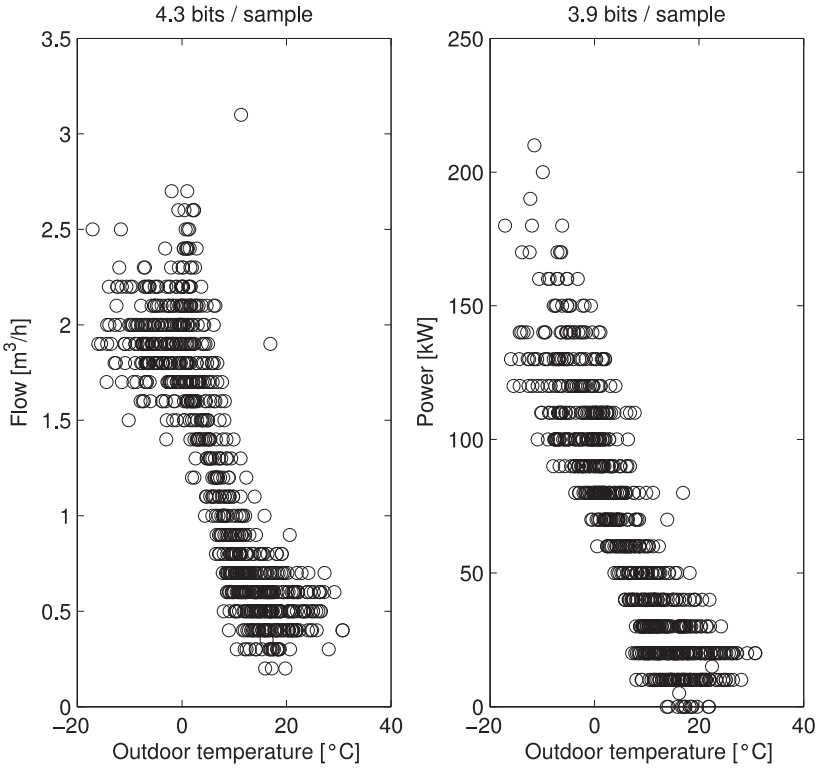


Figure 6.1. Data from an apartment building with abnormal quantization of the flow and power data, which results in a low information content of the hourly flow and power samples (4.3 bits per sample and 3.9 bits per sample, respectively).

6.1 Ranking with entropy

In information theory, *entropy* is a measure of the expected information contained in a message, for example in terms of *bits* (Murphy, 2012, Section 2.8). Claude Shannon introduced the information entropy concept in 1948 in his seminal mathematical theory of communication. The concepts of information and entropy are fundamentally related to *uncertainty*, the average unpredictability of a random variable. Entropy plays an important role in a wide range of applications and research fields, and it is fundamentally related to the thermodynamics of physical systems by the generalized Jarzynski equality. That being said, the description of entropy that is included here is pragmatic and focused on the use of entropy for the detection of abnormal energy metering data with a ranking procedure, which is similar to that proposed for outlier detection.

In order to appreciate the basic concept it is instructive to consider a typical textbook example of entropy; the coin-tossing example. When tossing a fair coin that has equal probabilities of coming to rest with either side of the coin facing upwards, the average entropy of each coin-toss experiment is one bit. There are two possible outcomes with equal probabilities, which means that the information that we gain about the state of the coin in one experiment is $\log_2 2 = 1$ bit. When tossing two fair coins simultaneously there are four possible outcomes of the experiment, so the information gained about the state of the two coins in one experiment is $\log_2 4 = 2$ bits. This idea generalizes to an experiment with n fair coins, which provides n bits of information, and to digital representations of information that are based on binary states (bits). In the context of digital technology it is often the inverse relation between the number of possible states, m, and the number of independent bits, n, that is of interest, $m = 2^n$, because it determines how many bits that are needed to represent a certain number of integer values. For example, $2^8 = 256$ is the number integer values that can be represented by eight bits (one byte).

Now consider what would happen if the coin is unfair, which means that it is more likely to come to rest with one particular side facing upwards. In that case we can predict the most frequent result and be right more often than we are wrong, which means that the information that we gain from each coin-toss experiment is less than one bit. In the extreme case that one particular side of the coin is always facing upwards the information that we gain from an experiment is zero because the outcome of the experiment is pre-determined (formally, the uncertainty about the outcome decreases with the number of experiments made and has a limit of zero). The key point is that the entropy is a measure of the average uncertainty about the value of a variable.

If a variable that represents a non-stationary macroscopic physical quantity like flow, temperature or power has high precision we expect that the variable should assume many different values over time and that the information entropy (uncertainty) of the variable should be high. In contrast, if the variable is stationary, or assumes a few discrete levels the uncertainty and entropy of the variable is low.

Therefore, entropy can be used as a measure for the detection of variables with abnormal variations. Both exceptionally low and high entropies can be of interest. For example, high entropies can result from noise and low entropies can result from poor precision or stationary signals. Formally, the entropy, H, is defined as the sum over probabilities, $p(x_i)$, of the different possible states, x_i, of a variable

$$H = -\sum_{i=1}^{m} p(x_i) \log_2 p(x_i). \tag{32}$$

The unit of the entropy is one bit when a base-2 logarithm is used. An implementation of this function with name `entropy` is included in the appendix. Next we use this function to demonstrate how substations with abnormally quantized variables can be identified and we derive an estimate for the magnitude of the quantization error.

6.2 Test results

We calculate the entropy for the hourly primary supply and return temperatures, the flow and the power of the 996 substations in the test set, see Figure 6.2.

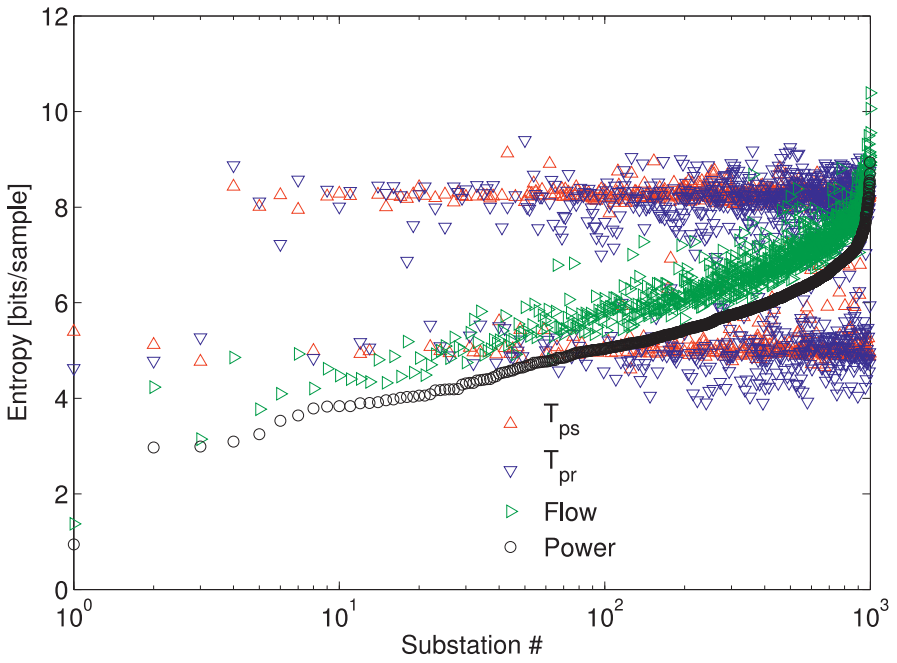


Figure 6.2. Entropy of hourly temperature, flow and power values for the population of 996 substations. The calculation is based on one year of data. Low entropies of the flow and power indicate that these substations may have oversized or misconfigured instrumentation, which results in abnormal quantization of the data.

The substations are ranked in order of increasing entropy of the power. The effective precisions of the temperature measurements are about 5 or 8 bits and vary between substations. There is an evident similarity between the entropies of the power and

flow data for different substations; Substations with low entropies of the power typically have associated low entropies of the flow. This is to be expected because a quantization of the flow variable affects the power calculated by the energy meter.

In Figure 6.3 we display the power profiles of the four substations in the test set that have the lowest entropies of the power variable. The quantization of the power is evident in all four cases and most likely depends on rounding of the data. Next we describe how the quantization error can be estimated in terms of the entropy.

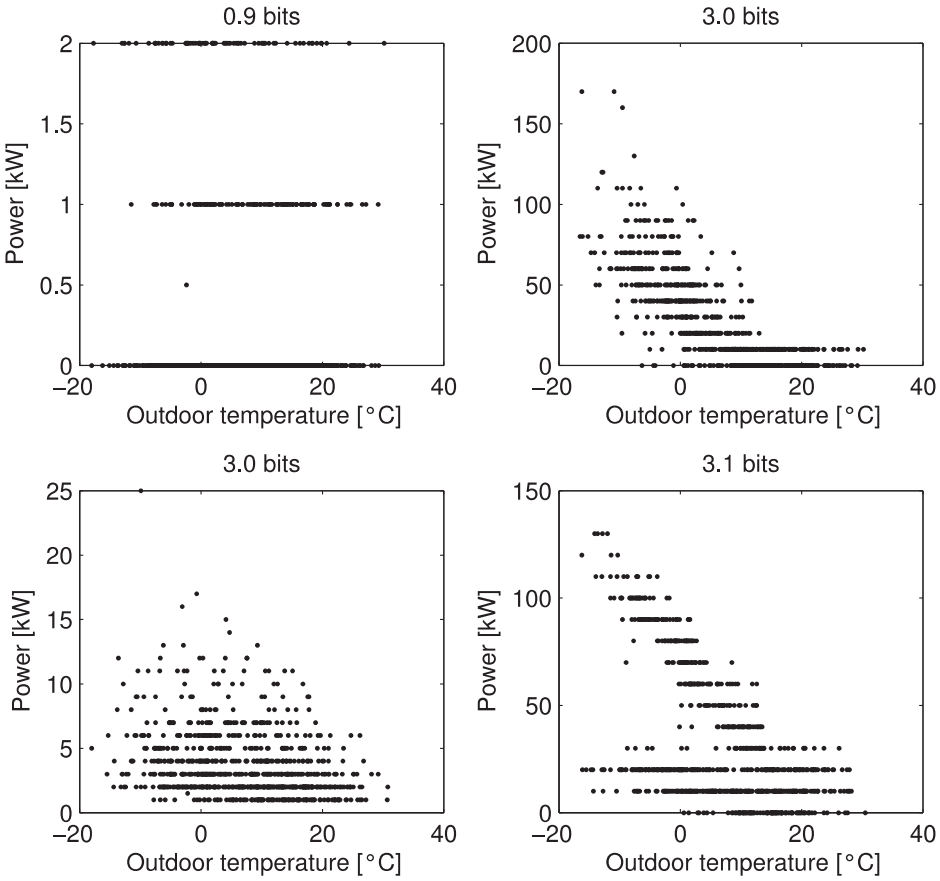


Figure 6.3. The four substations among the 996 substations in the test set with the lowest entropies of the power. For clarity, only 10% of the one-year dataset used to calculate the entropy is displayed.

6.3 Entropy and quantization error

The minimum entropy of the power that can be allowed if the maximum quantization error is to be kept below some limit can be estimated in the following way. Assuming that the quantized levels of the power are approximately equally spaced the step size between levels is about

$$\Delta P \approx \frac{P_{max}-P_{min}}{2^H}, \quad (33)$$

where H is the entropy and $P_{min}\sim 0$ is a reasonable approximation. Using this relation and the condition that $\Delta P \leq \alpha P_{max}$ the following limit is obtained for the entropy

$$\mathrm{H} \geq -\log_2 \alpha. \quad (34)$$

For example, if we allow a maximum quantization error of $\alpha = 10\%$ the entropy of the power should be higher than $-\log_2 0.1 \approx 3.3$ bits ($2^{3.3} \approx 10$). This implies that the quantization errors for the four substations that are illustrated in Figure 6.3 are above or near the 10% level. In the population of 996 substations there are 5 substations with power entropies below $-\log_2 0.1$ bits, 30 substations with entropies below $-\log_2 0.05$ bits and 169 substations with entropies below $-\log_2 0.025$ bits. Note that a low entropy does not necessarily imply that the energy communicated by the energy meter is incorrect because it can be a matter of misconfiguration or rounding in the communication between the energy meter and the data management system. Anyway, the detection of abnormal quantization is motivated if the data is to be used for analysis purposes and services.

6.4 Implementation

Entropy-based ranking can be used to identify substations with potential measurement precision problems, instrumentation faults or misconfigurations that result in abnormal quantization or stationary values of variables. This can be implemented in a similar way as the outlier and drift detection methods, by adding a column for the entropy of the power (possibly also the entropies of other variables) to the sortable list of substations. The entropy could either be calculated once for a specific reference time period, or it can be calculated online as illustrated in Figure 6.4.

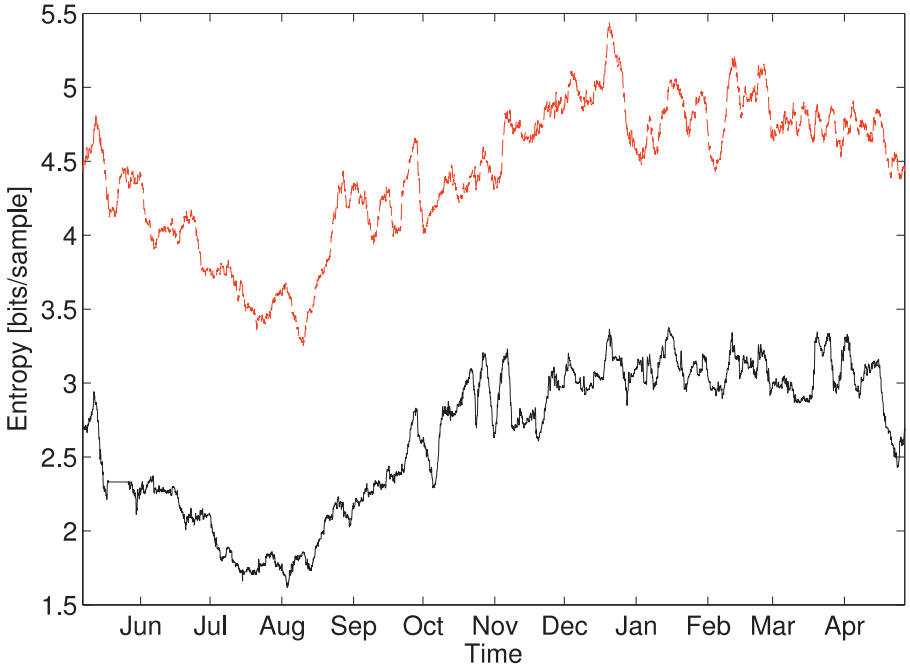


Figure 6.4. Weekly moving entropies of the power data for two different substations. One substation (solid line) is illustrated in Figure 6.1 and has evident quantization levels in the power profile, which results in a relatively low entropy. The other substation (dashed line) has higher entropy and no visible quantization levels in the power profile, see Figure 2.1.

7 DISCUSSION

This report deals with probabilistic methods and heuristics for automated detection and ranking of anomalies (potential faults), which can be implemented in existing district-energy data management systems. The development of such methods is motivated by difficulties experienced by utilities to identify faults in large-scale district energy systems and to process the high number of false alarms that can result when simplistic methods for fault detection are used. The problem to detect faults is important for several reasons; Faults that affect billing need to be avoided; Energy market regulations tend to become more restrictive with time in terms of accuracy, system efficiency and environmental effects; There is a growing interest to exploit the energy metering data for system optimization and development of new information services. Instrumentation faults are not uncommon in this type of system and can for example result in outliers, long-term drift and loss of information (see Chapter 1 for further information and examples). The methods that are commonly used for load analysis and fault detection are simplistic and range from degree-day estimates to basic statistical modelling and limit checking of daily or monthly average quantities. The estimated cost of faults that remain undetected for longer periods of time is substantial, and faults that lead to incorrect billing and information can affect customer relations and trust.

We focus on hourly energy metering data in this work, which is used by some district energy utilities and tend to become more common with time as the technology and data management software are upgraded. Hourly energy metering data offers significant advantages compared to daily or monthly averages. In particular, hourly values enable monitoring and analysis of intraday load cycles, improved system optimization strategies and customer information services. Intraday load cycles are directly related to peak loads, which can be costly since specialised peak-production plants that operate with fossil fuels are used. Therefore, monitoring and control of intraday cycles are important to enable improved system efficiency and a reduction of primary resource use, for example via price models that take intraday cycles and peak loads into account. Hourly data also offers improved sensitivity to faults that result in outliers, which can be difficult to detect when daily or monthly averages are considered. Note that the motivation for using hourly energy metering data comes from the effects of subsystems like ventilation systems, complementary heat sources and human behaviour, which can affect the average thermal power significantly at a timescale of one hour. The use of high-resolution data implies that more energy metering data is accumulated, which calls for efficient and automated methods for anomaly detection. The methods that are discussed in this report can in principle be applied also with daily energy metering data, except for the analysis of intraday cycles that is discussed in Section 3.1. If daily energy metering data is considered that method should be replaced with a method for analysis of intraweek cycles, for example the method that is described by Seem (2005, 2007) or Li et al. (2010). It

should be noted that the use of hourly district energy data is not unproblematic because the typical mix of hourly averages and instant samples in the management systems results in inconsistencies that can be significant, see Figure 1.2. This needs to be considered when designing methods for fault detection and diagnosis.

The methods that are described in this report are summarized in Table 7.1. Heuristic limit checking tests are discussed in Section 2.1 and are successfully used by some utilities. For example, some of these tests are implemented in the commercial energy data management software GENERIS (by Enoro AB) and Elin (by Powel Energy Management AB). Limit-checking methods are straightforward to implement and enable detection of some common faults. Therefore, we recommended that these methods should be implemented. In Chapter 3 we describe various probabilistic models of hourly energy meter data. The methods that are described in Chapters 4-6 for outlier detection, detection of long-term drift and detection of abnormal quantization are based on residual analysis of the probabilistic models. The "basic model" that is listed in Table 7.1 is the simplified method for outlier detection that is described in Section 2.2, which does not require fitting of probabilistic models to historical data. The basic model is less accurate than the methods that are described in the subsequent chapters; in particular it is insensitive to long-term drift and faults that do not result in short-term outliers in the data. Outlier detection and scoring is useful for rapid detection of abnormal data, which is essential if faults are to be detected before customers potentially notice the effects on the control system or bills. Also, outlier detection should be implemented if services that depend on hourly data are developed and exposed to customers because outliers can often be spotted.

Table 7.1. List of methods (columns) versus quantities considered for anomaly detection (rows). Symbols indicate whether a method is applicable (+) or not (-). These methods are described in Chapters 2-6 of this report.

	Limit-checking	**Basic model**	**Cluster analysis**	**Regression modelling**	**Outlier detection**	**Drift detection**	**Quantization analysis**
P	+	+	+	+	+	+	+
V	+	+	+	+	+	+	+
T_{ps}	+	+	–	+	+	+	+
T_{pr}	+	–	–	–	–	–	+

Scoring and ranking of anomalies is a central concept that is discussed in this report. We find that statistical hypothesis tests are error prone because of the complex and varying dynamics of the many buildings in a district energy system, in combination with the low sampling frequencies that typically are used in energy data management systems. The residuals of the models that we consider in this work have varying probability distributions, which sometimes are similar to a normal distribution, sometimes are similar to a Weibull distribution, sometimes are multimodal and

sometimes seem to result from a combination of different distributions for the mode and tails. Therefore, without going into detailed modelling of individual buildings it does not make sense to define definite decision limits for fault detection beyond the basic heuristic limits that are defined in Section 2.1. Instead, we propose that abnormal data is scored according to the deviation from the expectation and that a ranking procedure is used to identify outstanding substations for further investigation. The effective thresholds that determine whether further investigation is motivated will in practise depend on the resources that are available and the experience of the operators. Scores can be defined in an absolute or relative way, which makes a difference when simultaneously considering a population of substations with both high and low average loads. For example, a minor outlier in the data from a high-load substation can be scored higher than a major outlier in the data from a low-load substation when an absolute score is used, but the consequences of the outliers in terms of customer relations and trust can be more significant for the substation with low load, low absolute scores and relatively low economic risk. Scoring of outliers in building energy data is also discussed by Seem (2007).

We study the proposed methods using hourly data from a population of 996 district heating substations. We have developed and used a Matlab implementation of the methods that are presented in this report, which enables automated analysis of data from several thousands of district energy substations. Sample code of key functions needed to implement the methods is provided in the appendix. This does not mean that the methods that are presented in this report are ready to use “as is”. Rather, they should be considered as proof of concept. It remains to learn which methods that will prove useful and cost-effective in a full-scale implementation. A real-world implementation also needs to account for eventual exceptions such as holidays, which is a consequence of the strict analysis of intraweek cycles that is described in Section 3.1. Seem (2005) proposes a different approach for the analysis of intraweek cycles, which is less sensitive to holidays because each day is classified in terms of a few “day types” rather than weekdays.

At the end of this project we learned that Göteborg Energi AB uses piecewise linear regression to model the relationships between outdoor temperature and power, and the outdoor temperature and flow. The implementation, named Kasper, is developed since 2002 with support from two Master’s Thesis students (Munoz, 2006; Lindquist, 2010) and Professor Anders Odén at the Chalmers University of Technology. Also, a similar approach is under development for wind power (Forsman, 2011). In Kasper, the regression models have three fixed breakpoints and are fitted to daily average values for a given time period and subset of weekdays (high versus low load etc.). Substations are ranked according to the size of the regression model residuals, and a one-year moving average is used to identify drift. The regression models are also used for imputation of missing energy meter data, and as a visualization tool in the communication with customers. This means that two groups have independently made similar assumptions and come to similar

conclusions concerning the modelling approach and use of ranking heuristics. There are also some differences between the two approaches. We have considered hourly data and intraday cycles in addition to the intraweek cycles / day types, which is necessary when hourly values are considered, and we have adopted the approach described by Seem (2007) to identify and score potential outliers in the model residuals. Also, we describe an approach based on CUSUM control charts for detection of drift, and we introduce the concept of Shannon entropy for detection of oversized and misconfigured instrumentation. The implementation of piecewise linear regression is more refined in Kasper compared to our prototype implementation, and demonstrates that fault detection with piecewise regression models and residual analysis is applicable and useful in practise.

In principle the methods that are discussed in this report may be applicable also to district cooling data, but we do not study that possibility here. The lower primary temperature difference in district cooling applications makes the fault detection problem a more delicate one. Therefore, we propose that the methods that are described here should first be implemented and evaluated in district heating systems.

7.1 Directions for further work

The limited information, low sampling rate and combination of instantaneous samples and averaged quantities that are commonly accessible in district energy management systems render the situation non-ideal for automated fault detection and information quality assessment. An optimal approach requires a technology and culture transition towards substations with holistic monitoring and control systems. That would enable integration of standard methods for fault detection and diagnosis, for example dynamical models and state estimation techniques. A step of that magnitude can be motivated if system optimization aspects and environmental effects are included in the picture. For example, holistic monitoring and control systems can be used to improve system efficiency (Delsing et al., 2009) and new price models can reduce the daily peak demands without compromising comfort (van Deventer et al., 2011).

Some modern energy meters have integrated basic functionality for fault detection. An intermediate step is to develop more efficient methods for fault detection and diagnosis that either can be implemented in the energy meters, or the interfaces to the data management systems. By integrating the fault detection and diagnosis methods near the energy meter it is possible to access and process high-resolution data, at least for the primary flow and temperatures. The communication bottleneck that often exist between the energy meters and the data management systems needs to be respected and managed if this approach is to be useful for the industry, and it must be easy to configure the fault detection and diagnosis methods in a large-scale system.

There is also room for improvement of the probabilistic methods that are described in this work in order to increase the sensitivity and robustness of the anomaly detection approach. In particular, the understanding of the probability distributions of the model residuals is incomplete and can be further investigated in

order to understand the underlying dynamics and derive improved probabilistic models. The analysis of intraday cycles can possibly be improved by considering an approach similar to that used by Li et al. (2010), or multivariate regression. Further work is needed to understand how faults affecting the primary return temperature can be detected directly, without reference to the power. This problem can be solved in principle by monitoring of the heat exchangers (Isermann, 2011), but it is difficult given the information that is commonly available in the management systems.

The lack of a well-defined dataset makes the development and evaluation of methods for fault detection challenging, and the fact that historical energy metering data includes abnormal data is often ignored in the literature. Therefore, it would be useful if a reference dataset could be assembled for researchers and developers to use, which should include several years of data for a representative number of substations that have been confirmed as, respectively, faulty and non-faulty. Faults should be of varying type and need to be described in the dataset.

8 CONCLUSIONS

Faults are common in district energy systems due to the high number of substations, which include several instrumentation components like mechanical flow meters, temperature sensors, cabling and electronic devices. Also, the instrumentation is designed for low cost and billing, not for redundant measurement and automated detection of faults. The methods that are commonly used for fault detection are simplistic and implemented at the energy data management level. A deluge of false alarms that is difficult and costly to analyse can result from the application of such methods. In summary, we observe that:

- Abnormal measurements and instrumentation faults are detected for several percent of the substations in a district heating system on an annual basis.
- The cost of faults can be high, for example of the order 50 000 € per year in the case of a drifting flow meter for a large customer.
- The methods that are commonly applied for fault detection are simplistic.
- Some utilities manually handle thousands of false alarms per month, which is a costly, counter-productive and error-prone task.
- There is a need for improved fault-detection methods.
- There is a growing interest among the utilities to develop services and system optimization strategies that depend on high-resolution data, which requires more efficient methods for fault detection and quality assessment of the data.

Therefore, development of improved approaches and methods for fault detection in district heating systems is motivated. In this report we present methods for detection and ranking of anomalies, which can be used to automatically identify potential faults in district energy metering data. In particular, we present methods for:

- Regression modeling of variable relationships.
- Analysis of intraday and intraweek cycles in the data.
- Outlier detection and ranking.
- Drift detection and ranking.
- Detection of abnormal quantization (design faults, misconfiguration etc.).

We study the proposed methods with data from a population of 996 substations and demonstrate that we can identify substations with documented faults, unknown faults and abnormal characteristics. In this population about 5% of the substations are abnormal and about half of these have been confirmed as faulty. The methods are selected and designed with automated use and applicability in mind, which is crucial to enable cost-efficient analysis of the data. The proposed methods need to be implemented by utilities in a full-scale district energy management system before the effects on the fault detection rate and cost efficiency can be properly evaluated.

9 ACKNOWLEDGEMENTS

The Swedish District Heating Association, the Swedish Energy Agency, Enoro AB and Powel Energy Management AB funded this project. Several district energy companies have supplied data and valuable working knowledge to this project, which is an important basis for the work and results that are summarized in this report. We thank the industrial representatives that participated in the reference group of the project for their efforts and helpful advice. In particular, we thank Robert Eklund for contributing several useful ideas. This project has been a valuable learning experience that has stimulated ideas for further research, in particular for anomaly detection in complex systems where models and variable relationships are unknown. We hope that the methods that are proposed here will be implemented in an energy meter data management system and that they will prove useful for the district energy industry.

10 REFERENCES

Armstrong, P. R., Leeb, S. B. and Norford, L. K., *Control with Building Mass - Part I: Thermal Response Model*, ASHRAE Transactions 112 (2006).

Bergquist, T., Ahnlund, J., Johansson, B., Gårdman L. and Råberg, M., *Alarm reduction with correlation analysis* (in Swedish), Värmeforsk Service, Stockholm, 60p (2004).

Berrebi, J., *Self-Diagnosis Techniques and their applications to error reduction for ultrasonic flow measurement*, Ph.D. thesis, Luleå University of Technology, Sweden, 130p (2004).

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C., *Time Series Analysis: Forecasting and Control*, Wiley Series in Probability and Statistics, New Jersey, 746p (2008).

Carlander, C., *Installation effects and self diagnostics for ultrasonic flow measurement*, Ph.D. thesis, Luleå University of Technology, Sweden (2001).

Delsing, J. and Svensson, B., *En ny metod för funktionsdiagnos och felsökning av fjärrvärmecentraler utifrån värmemängdsdata*, Technical Report 2001:15, Luleå University of Technology, Sweden (2001).

Delsing, J., van Deventer, J. and Gustafsson, J., *Integrerad energimätning och reglering i en fjärrvärmecentral*, Technical Report, Fjärrsyn, Svensk Fjärrvärme, 36p (2009).

Forsman, N., *An analytical tool for the evaluation of wind power generation,* Master's Thesis, Department of Energy and Environment, Chalmers University of Technology, Sweden (2011).

Frederiksen, S. and Werner, S., *Fjärrvärme teori, teknik och funktion*, Studentlitteratur, Sweden, 440p (2001).

Friedman, J. H., *Multivariate Adaptive Regression Splines*, Annals of Statistics 19, pp. 1–67 (1991).

Kiluk, S., *Algorithmic acquisition of diagnostic patterns in district heating billing system*, Applied Energy 91, pp. 146–155 (2012).

Iglewicz, B. and Hoaglin, D., *Volume 16: How to Detect and Handle Outliers*, The ASQC Basic References in Quality Control: Statistical Techniques, (1993).

Isermann, R., *Model-based fault-detection and diagnosis – status and applications*, Annual Reviews in Control 29, pp. 71–85 (2005).

Isermann, R., *Fault-Diagnosis Systems – An Introduction from Fault Detection to Fault Tolerance*, Springer, Berlin Heidelberg, 475p (2006).

Isermann, R., *Fault-Diagnosis Applications*, Springer, Berlin Heidelberg, 354p (2011).

Jiménez, M. J. and Madsen H., *Models for describing the thermal characteristics of building components*, Building and Environment 43, pp. 152–162 (2008).

Jiménez, M. J., Madsen, H. and Andersen, K.K., *Identification of the main thermal characteristics of building components using MATLAB*, Building and Environment 43, pp. 170-180 (2008).

Johansson, A., *Fault detection of hourly measurements in district heat and electricity consumption*, M.Sc. thesis, Department of Electrical Engineering, Linköping University, 60p (2005).

Jomni, Y., *Improving heat measurement accuracy in district heating substations*, Licentiate Thesis, Luleå University of Technology, Sweden, 127p (2004).

Jota, P., Silva, V. and Jota, F., *Building load management using cluster and statistical analyses*, Electrical Power and Energy Systems 33, 1498–1505 (2011).

Katipamula, S. and Brambley, M. R., *Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems – A Review, Part I*, International Journal of HVAC&R Research 11, No. 1, pp. 3-25 (2005).

Katipamula, S. and Brambley, M. R., *Methods for Fault Detection, Diagnostics, and Prognostics for Building Systems – A Review, Part II*, International Journal of HVAC&R Research 11, No. 2, pp. 169-187 (2005).

Li, X., Bowers, C. P. and Schnier, T., *Classification of Energy Consumption in Buildings With Outlier Detection*, IEEE Transactions on Industrial Electronics 57, No.11, pp. 3639-3644 (2010).

Lindquist, P., *Verktyg för utvärdering av energieffektiviseringar, baserat på effektsignaturanalyser*, M.Sc. thesis, Department of Computer Science and Engineering, Chalmers University of Technology, 45p (2010).

Munoz, M., *Kvalitetsstudie för fjärrvärmemätare*, M.Sc. thesis, Department of Mathematical Sciences, Chalmers University of Technology, 52p (2006).

Murphy, K. P., *Machine Learning: a Probabilistic Perspective*, The MIT Press, 1104p (2012); http://www.cs.ubc.ca/~murphyk/MLbook/index.html.

NIST / SEMATECH, *e-Handbook of Statistical Methods*, http://www.itl.nist.gov/div898/handbook/, last updated at 2012-04-01.

Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W. (editors), *NIST Handbook of Mathematical Functions*, Cambridge University Press, New York, NY (2010); Online companion: *NIST Digital Library of Mathematical Functions*, http://dlmf.nist.gov/, release 1.0.5 at 2012-10-01.

Pakanen, J., Hyvärinen, J., Kuismin, J. and Ahonen, M., *Fault diagnosis methods for district heating substations*, Research Notes 1780, Technical Research Centre of Finland (VTT), Espoo, 70p (1996).

Reference group, industrial expert reference group of the project "Validation of data for energy metering" (Validering av data för energimätning) within the Fjärrsyn R&D program, 2011 - 2013. Contact Svensk Fjärrvärme AB for further information.

Rosner, B., *Percentage points for a generalized ESD many-outlier procedure*, Technometrics 25, pp. 165-172 (1983).

Ryu, J. H., Wan, H. and Kim, S., *Optimal design of a CUSUM chart for a mean shift of unknown size*, Journal of Quality Technology 42, pp. 311-326 (2010).

Sandin, F., Gustafsson, J., Eklund, R. and Delsing, J., *Basic methods for automated fault detection and energy data validation in existing district heating systems*, 13th International Symposium on District Heating and Cooling, September 3–4, Copenhagen, Denmark, pp. 183-190 (2012).

Seem, J. E., *Pattern recognition algorithm for determining days of the week with similar energy consumption profiles*, Energy and Buildings 37, pp. 127-139 (2005).

Seem, J. E., *Using intelligent data analysis to detect abnormal energy consumption in buildings*, Energy and Buildings 39, pp. 52–58 (2007).

Simonoff, J. S., *A comparison of robust methods and detection of outliers techniques when estimating a location parameter*, Communications in Statistics - Theory and Methods 13, pp. 813-842 (1984).

Svensson, B., *In-situ test methods in district heating systems*, Licentiate thesis, Lulea University of Technology, Sweden (1996).

Svensk Fjärrvärme, *Om fjärrvärme*, http://svenskfjarrvarme.se; http://svenskfjarrvarme.se/Fjarrvarme/ (accessed May 7, 2013).

Taylor, J. W. and McSharry, P. E., *Short-Term Load Forecasting Methods: An Evaluation Based on European Data*, IEEE Transactions on Power Systems 22, pp. 2213-2219 (2007).

Tukey, J. W., *Exploratory data analysis*, Reading, MA 231, Addison-Wesley (1977).

van Deventer, J., Gustafsson, J. and Delsing, J., *Controlling district heating load through prices*, 2011 IEEE International Systems Conference, Montreal, Canada, pp. 461-465 (2011).

Woodall, W. H. and Adams, B. M., *The Statistical Design of CUSUM Charts*, Quality Engineering 5, pp. 559-570 (1993).

Yliniemi, K., *Fault detection in district heating substations*, Licentiate thesis, Luleå University of Technology, Sweden, 95p (2005).

APPENDIX 1 – Linear regression software

Linear regression is common in fields like statistics, machine learning, data mining and explorative data analysis. Therefore, there are many implementations of algorithms for linear regression, which often include aspects like robustness towards outliers in the data and adaptive optimization of the position and number of segments (Friedman, 1991; Murphy, 2012, Chapter 16). Extensions to nonlinear functions of the data are also common. Some examples of software packages and libraries that can be used for linear regression are listed below.

Free and open source

- R Programming Language (earth, mda and polspline packages); http://r-project.org.
- Orange for Python (earth package); http://orange.biolab.si.
- ARESLab for Octave/Matlab; http://www.cs.rtu.lv/jekabsons/regression.html.
- SPLINEFIT; http://www.mathworks.com/matlabcentral/fileexchange/13812-splinefit

Commercial

- Matlab Curve Fitting Toolbox; http://www.mathworks.se/discovery/data-fitting.html
- MARS from Salford Systems; http://www.salford-systems.com.
- STATISTICA Data Miner from StatSoft; http://www.statsoft.com.

In this work we use piecewise linear relationships because such functions represents the empirical data sufficiently well and a simple approach is motivated by the need to make tens of thousands of regression models in an automated fashion. A piecewise linear approach is also motivated by the common use of piecewise linear functions in the control system of district heating substations. When generating many regression models automatically for a large population of district energy substations the extra degrees of freedom associated with higher-order polynomials tend to make the results less reliable in the sense that some regression models are unnatural. ARESLab and SPLINEFIT have been used to produce the figures and results presented in this report. ARESLab supports adaptive regression. The optimization of segments is done in two steps, first by adding breakpoints to reduce the variance, then by pruning the breakpoints so that a reasonable trade-off between accurate fit and model complexity is achieved. Adaptive methods are useful for manual modelling, but are tricky to implement in a reliable way for automated generation of many models. Unless there is a good reason to use adaptive regression, we recommend that a fixed number of about five to ten segments is used. A formal introduction to robust linear regression can be found in Murphy (2012).

APPENDIX 2 – Code samples

The core functions used to calculate the figures and results presented in this report are listed here. Software for linear regression is listed in Appendix 1.

Basic online test for outliers

```
function [i,Z] = basic_test(s)
% Basic online test and ranking of outliers.
%   This method is a simplified form of the method described by
%   John E. Seem (2007) for detection of outliers in the power,
%   which can be generalized for detection of outliers also in
%   the flow and supply temperature. One-week moving averaging
%   is used to estimate and subtract the trend caused by varying
%   outdoor temperature and seasons. A one-week moving average is
%   used to average out intraday and intraweek cycles.
%   This results in a simple online algorithm that can be used for
%   automated detection and ranking of outliers in the power,
%   flow and supply temperature (see *** below), which does
%   not involve analysis of long-term historical records of
%   data and is straightforward to implement.
%
%   s      Substation data record.
%   i      Indices of outliers.
%   Z      Modified Z scores of outliers (higher is worse).
alpha = 0.05; % 95% confidence (ideal)
nout = 100;   % If outliers are found we fetch the top-100
nzdisp = 10;  % Number of top-Z scores to display
t = s.time;   % Testing can be limited to data for the last few weeks
y = s.power;  % Replace with s.flow to validate the flow, or            ***
              % replace with T_ps or ΔT_ps to validate the supply temperature,
              % where ΔT_ps is the supply-temperature difference between
              % two different substations with highly correlated
              % supply temperatures, see "correlation analysis" below.

% Temperature de-trending with weekly moving average
ym = moving(y, 7*24); % 7*24 hourly values
yy = y - ym;

% GESD outlier detection
[i,~,~] = gesd(yy, nout, alpha);
j = true(1,length(y));
j(i) = false;
% Modified Z scores
Z = yy(i) / std(yy(j));

% Plot outliers and display Z scores
semilogy(t,y,'k-',t(i),y(i),'rx');
for k=1:min([nzdisp length(Z)])
    lbl = sprintf('%1.1f',Z(k));
    text(t(i(k))+0.01*range(xlim), y(i(k))+0.01*range(ylim), lbl);
    datetick;
    xlabel('Time');
    ylabel('Power [kW]');
    % ylabel('Flow [m^3/h]');           % If the flow is considered     ***
    % ylabel('\Delta T_{ps} [°C]');     % or the supply temperature
end
```

GESD test for outliers

```
function [i,R,lambda] = gesd(x,r,alpha)
% GESD test for outliers.
%
%   x      Data vector.
%   r      Maximum number of outliers to search for.
%   alpha  Significance level (for example 0.05 for 95% cf).
%   i      Indices of outliers in x.
%   R      R statistic.
%   lambda Critical values.
%
%   For details, see ref. Seem 2007 and
%   http://www.itl.nist.gov/div898/handbook/eda/section3/eda35h3.htm

    % Calculate test statistic and critical values
    R = zeros(1,r);
    ivec = zeros(1,r);
    lambda = zeros(1,r);
    n = length(x);
    xtemp = x;
    itemp = 1:length(x);

    for i=1:r
        m = mean(xtemp);
        s = std(xtemp);
        [R(i),j] = max(abs(xtemp-m)/s);
        ivec(i) = itemp(j);
        xtemp(j) = []; % delete R(i)
        itemp(j) = [];
        p = 1 - alpha/(2*(n-i+1));
        t = tinv(p,n-i-1); % Inverse of Student's CDF
        lambda(i) = (n-i)*t/sqrt((n-i+1)*(n-i-1+t*t));
    end

    % Number of outliers is determined by highest i so that R(i) > lambda(i)
    i = max(find(R>lambda));
    i = ivec(1:i);

end
```

Standardized power

```
function ystd = std_power(s,tmax)
% Calculate standardized power versus the different weekdays and time of day.
%    The result is used as input for the cluster analysis and calculation of
%    weekly schedules.
%
%   s      Substation data record.
%   tmax   Outdoor temperature threshold (cycles are more evident at low
%          outdoor temperature).
%   ystd   A 7x24 cell array of vectors with standardized power.

PLOT = 0;

x = s.t_outdoor;
y = s.power;

w = weekday(s.time) - 1; % Sun Mon Tue Wed Thu Fri Sat
w(w==0) = 7; % Mon Tue Wed Thu Fri Sat Sun
h = hour(s.time);

% Calculate standardized y for each weekday and hour
ystd = cell(7,24);
tmin = floor(min(s.t_outdoor));
for t=tmin:(tmax-1)

    % 1°C binning, remove outdoor temperature dependence
    i = (s.t_outdoor >= t & s.t_outdoor < (t+1));
    yi = y(i);

    % Calculate standardized value
    yi = yi - mean(yi);
    yi = yi ./ std(yi);

    % Extract weekdays and time of day
    wi = w(i);
    hi = h(i);

    % Append values to tuple
    for j=1:length(yi)
        ystd{wi(j),hi(j)+1} = [ystd{wi(j),hi(j)+1} yi(j)];
    end
end

% Plot mean[ystd] and std[ystd]
if PLOT
    hold off;
    pm = zeros(7,24);
    sm = zeros(7,24);
    style = {'k-','r-','g-','b-','m-','k--','r--'};
    for i=1:7 % weekdays
        for j=1:24 % time of day
            pm(i,j) = mean(ystd{i,j});
            sm(i,j) = std(ystd{i,j});
        end
        errorbar(0:23,pm(i,:),sm(i,:),style{i});
        hold on;
    end
    hold off;
    legend('Mon','Tue','Wed','Thu','Fri','Sat','Sun');
    xlabel('Time');
    ylabel('P_s');
end

end
```

Bimodality coefficient

```
function bc = bimodality(x)
% Bimodality coefficient (BC) of distribution in the vector x.
%
%   Formula suggested by Warren Sarle.
%   The BC is 1 for a Bernoulli distribution (maximum bimodality),
%   1/3 for a normal distribution, and near zero for heavy-tailed
%   distributions. The statistic of the BC is unknown. Another simple
%   measure for bimodality is negative kurtosis, but that measure has
%   limitations that motivated the invention of the BC. There are
%   several tests for bimodality proposed in the literature, but none
%   that offers a universal solution. Empirical results indicate that
%   the BC is a useful and simple indicator for bimodality in district
%   energy power data. As a rule of thumb, a value higher than about
%   0.65 corresponds to bi- or multimodal distributions, while lower BC
%   values indicate that the power distribution is practically unimodal.

    bc = (skewness(x)^2+1) / kurtosis(x);

end
```

Entropy

```
function s = entropy(x)
% Entropy of a set of numbers in bits.

    x = reshape(x,1,numel(x));

    % Count how many times each number occurs
    [~,n,~] = unique(sort(x));
    n = [n(1) diff(n)];

    % Calculate probabilities of numbers
    p = n ./ numel(x);

    % Calculate entropy (observe inner product)
    s = -p*log2(p)';

end
```

Cluster analysis

```
function schedule = power_schedule(s, tmax)
% Make weekly schedule by cluster analysis of standardized power.
%   Divide weekdays and time of day into three clusters
%   corresponding to low, intermediate/mixed and high power.
%   The returned schedule is a 7x24 matrix with -1 for low,
%   0 for intermediate/mixed and +1 for high power. Intermediate
%   should be interpreted as "both high and low" because that
%   class typically includes data points belonging to both
%   the low and high power clusters. When fitting regression
%   models, the mixed cluster should be excluded. In limit checking
%   and outlier detection the mixed-class data should be treated
%   with conservative limits (the lower limits of the low-power
%   cluster and the upper limits of the high-power cluster).
%   For other approaches to identify clusters, see Seem 2005
%   and Li et al 2010. The method proposed here is more simple
%   than the method outlined in Li 2010, and we consider
%   intraday cycles in addition to the intraweek cycles
%   (Seem and Li et al focus on the daily average power).
%
%   s         Substation data record.
%   tmax      Outdoor temperature threshold (cycles are more evident at low
%             outdoor temperature).
%   schedule  A 7x24 matrix of cluster identifiers:
%             -1    low power demand,
%              1    high power demand,
%              0    mixed, both high and low power demand.

% Toggle this to plot weekly schedule.
PLOT = 1;

% Toggle safe transitions. Inserts mixed class at
% transitions from low to high power and vice versa.
SAFE = 1;

% Calculate standardized power
pstd = std_power(s,tmax);

% Create a vector of mean standardized power for cluster analysis.
% This approach can be extended by including the variance of the
% standardized power as an indicator that (i,j) should be included
% in the "mixed" cluster.
pv = zeros(4,7*24);
for i=1:7
    for j=1:24
        pv(1,(i-1)*24+j) = i;
        pv(2,(i-1)*24+j) = j-1;
        pv(3,(i-1)*24+j) = mean(pstd{i,j});
    end
end

% Cluster analysis. Divide data in two clusters, then add
% a point for an intermediate cluster and repeat the analysis.
[~,c] = kmeans(pv(3,:)',2,'Start',prctile(pv(3,:),[10; 90]));
[ik,c] =
kmeans(pv(3,:)',3,'Start',[min(c);mean(c);max(c)],'EmptyAction','drop');
nclust = 3;
if sum(isnan(c))==0 % 3 clusters (low,mixed,high)
    imin = 1;
    imid = 2;
    imax = 3;
elseif sum(isnan(c))==1 % 2 clusters (low,high)
    if isnan(c(2))
        imin = 1;
        imax = 3;
    elseif isnan(c(1))
```

```
        imin = 2;
        imax = 3;
    else
        imin = 1;
        imax = 2;
    end
    imid = -1; % disable
    nclust = 2;
else
    error('Only one cluster, BC is too low?');
end

% Create schedule
schedule = zeros(7,24);
schedule(sub2ind(size(schedule),pv(1,ik'==imin),pv(2,ik'==imin)+1))=-1; % low
schedule(sub2ind(size(schedule),pv(1,ik'==imid),pv(2,ik'==imid)+1))=0; % mixed
schedule(sub2ind(size(schedule),pv(1,ik'==imax),pv(2,ik'==imax)+1))=1; % high

% Label transitions between high/low power as intermediate power.
% The idea is simple: scan through the whole week sequentially in
% time. If one particular hour is classified as 'high power' and
% the next hour as 'low power' (or vice versa) then both hours are
% re-classified as 'mixed' to reduce the risk of misclassification.
if SAFE == 1
    nclust = 3;
    ss = reshape(schedule',1,7*24);
    ss = [ss ss(1)]; % repeat first element to simplify indexing
    for i=1:7
        for j=1:24
            if ss((i-1)*24+j)*ss((i-1)*24+j+1) == -1
                schedule(i,j) = 0;
                if j<24
                    schedule(i,j+1) = 0;
                else
                    if i<7
                        schedule(i+1,1) = 0;
                    else
                        schedule(1,1) = 0;
                    end
                end
            end
        end
    end
end

% Plotting
if PLOT == 1
    hold off;
    [h,d] = find(schedule == -1); % low power
    plot(h,d,'bv');
    hold on
    [h,d] = find(schedule == 0); % mixed
    plot(h,d,'kd');
    [h,d] = find(schedule == 1); % high power
    plot(h,d,'r^');
    xlabel('Weekday');
    set(gca,'XTickLabel',{'Mon','Tue','Wed','Thu','Fri','Sat','Sun'})
    ylabel('Time of day');
    ylim([0 23]);
    if nclust==3
        legend('Low power','Mixed','High power');
    else
        legend('Low power','High power');
    end
end

end
```

CUSUM statistic

```
function [Sm,Sp] = cusum(x,e,k,a)
% Cumulative sum statistic for bias detection.
%
%   x sample values
%   e expectation values
%   k reference level
%   a normalization constant
%
%   See http://www.itl.nist.gov/div898/handbook/pmc/section3/pmc323.htm
%   and Woodall & Adams, 1993 for further information.

Sm = zeros(1,length(x));
Sp = zeros(1,length(x));

for i=1:length(x)
    j = max([1 i-1]);
    Sm(i) = max([0, Sm(j) - (x(i) - e(i) + k)/a]);
    Sp(i) = max([0, Sp(j) + (x(i) - e(i) - k)/a]);
end

end
```

Correlation analysis

```
function [id,cor,dist] = find_correlated(id,data,top)
% Identify substations with highly correlated supply temperatures.
%             Arguments
%             id      ID of substation.
%             data    Substation database.
%             top     Number of geographical neighbors to consider,
%                     a higher number is better (~100 is usually ok).
%             Return values
%             id      Array of substation IDs (most correlated first).
%             cor     Array of correlation coefficients (descending order).
%             dist    Array of geographical distances.
    cmax=-1e10;
    cor = zeros(1,top);

    % Find nearby substations. This is not necessary but is done
    % here to reduce processing time. The use of geographical data
    % can be avoided by including all substations in the network,
    % which result in better matches at the cost of processing time.
    % This function is defined below.
    [nearby,dist] = find_nearby(id,data,top);

    % Calculate correlation coefficients for each substation
    s = get_substation(id,data);
    for i=1:length(nearby)
        s2 = get_substation(nearby(i),data);
        % Align time sequences (skip head/tail and missing datapoints)
        [j,k] = timealign(s.year,s.month,s.day,s.hour,s.minute,
             s2.year,s2.month,s2.day,s2.hour,s2.minute);
        % Calculate correlation coefficient,
        % (OBS de-trending with first-order differences)
        c = corrcoef(diff(s.t_ps(j)), diff(s2.t_ps(k)));
        cor(i) = c(1,2);

    end

    % Sort substations using the correlation coefficients
    [cor,ind] = sort(cor,'descend');
    id = nearby(ind);
    dist = dist(ind);

end
```

```
function [id,d] = find_nearby(id,data,top)
% Identify substations that are geographically nearby.
%    This function is not necessary, but is used there to
%    speed up the search for high-correlation pairs
%    (the search is limited to a geographical neighborhood).
    d = zeros(1,n);
    n = length(data.pos_id);
    % Calculate distances to all other substations
    s = get_substation(id,data);
    for i=1:n
        x = data.pos_x(i);
        y = data.pos_y(i);
        d(i) = (x-s.pos_x)^2 + (y-s.pos_y)^2;
    end
    d = sqrt(d);

    % Sort distances
    [d,id] = sort(d);

    % Get ID numbers of substations
    id = data.pos_id(id);

    % Remove self reference
    d = d(id ~= s.id);
    id = id(id ~= s.id);

    % Remove remote substations and substations
    % without associated energy data
    inc = false(1,length(id));
    idlist = unique(data.id);
    for i=1:length(id)
        if ismember(id(i),idlist)
            inc(i) = true;
        end
        if sum(inc)==top
            break;
        end
    end
    id = id(inc)';
    d = d(inc);
end
```

FJÄRRSYN

Forskning som stärker fjärrvärme och fjärrkyla, uppmuntrar konkurrenskraftig affärs- och teknikutveckling och skapar resurseffektiva lösningar för framtidens hållbara energisystem. Kunskap från Fjärrsyn är till nytta för fjärrvärmebranschen, kunderna, miljön och samhället i stort. Programmet finansieras av Energimyndigheten tillsammans med fjärrvärmebranschen och omsätter cirka 19 miljoner kronor om året. Mer information finns på www.fjarrsyn.se

VALIDERING AV MÄTDATA

Nya regler och ny teknik innebär att förbrukningen av fjärrvärme och fjärrkyla mäts allt oftare. Men fjärrvärmeföretagen anser att det är svårt att upptäcka fel i stora energisystem och att det är en utmaning att hantera och bearbeta en ökad mängd data.

Det är viktigt att felen inte förblir oupptäckta eftersom det kan bli kostsamt, men också att fjärrvärmeföretaget förlorar i trovärdighet till exempel när kunder upptäcker fel och får felaktiga räkningar.

Här beskrivs hur avvikelser i energimätdata kan upptäckas med automatiska metoder och ett minimum av mänsklig inblandning. Målet har varit att kunna analysera stora mängder mätdata på ett kostnadseffektivt sätt. Tack vare de nya metoder som beskrivs här kan avvikande mätvärden identifieras och rangordnas så att fjärrvärmeföretaget kan koncentrera sig på att analysera de centraler mest är mest avvikande. Rapporten är skriven på engelska, men har en svensk sammanfattning

Svensk Fjärrvärme

Svensk Fjärrvärme • 101 53 Stockholm • Telefon 08-677 25 50 • Fax 08-677 25 55
Besöksadress: Olof Palmes gata 31, 6 tr. • E-post fjarrsyn@svenskfjarrvarme.se • www.fjarrsyn.se