QUANTICOL

A Quantitative Approach to Management and Design of Collective and Adaptive Behaviours

quantiCOL

http://www.quanticol.eu

D5.1

Data validation and requirements for case studies

Revision: 1.0; March 26, 2014

Author(s):
Mirco Tribastone (SOTON), Allan Clark (UEDIN), Nicolas Gast (EPFL), Stephen Gilmore (UEDIN), and Daniël Reijsbergen (UEDIN)

Due date of deliverable: Month 12 (March 2014)

Actual submission date: March 31, 2014

Nature: R. Dissemination level: PU

Funding Scheme: Small or medium scale focused research project (STREP)
Topic: ICT-2011 9.10: FET-Proactive 'Fundamentals of Collective Adaptive Systems' (FOCAS)
Project number: 600708

Coordinator: Jane Hillston (UEDIN)
e-mail: Jane.Hillston@ed.ac.uk
Fax: +44 131 651 1426

Part. no.	Participant organisation name	Acronym	Country
1 (Coord.)	University of Edinburgh	UEDIN	UK
2	Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie della Informazione "A. Faedo"	CNR	Italy
3	Ludwig-Maximilians-Universität München	LMU	Germany
4	Ecole Polytechnique Fédérale de Lausanne	EPFL	Switzerland
5	IMT Lucca	IMT	Italy
6	University of Southampton	SOTON	UK
7	Institut National de Recherche en Informatique et en Automatique	INRIA	France

COOPERATION

Executive Summary

This deliverable reports on initial data-gathering activities regarding the case studies on smart public transportation (bus networks and bike sharing systems) and smart grids. It offers a critical validation of the quality of the information that is publicly available, and an analysis of the real data offered to us by project collaborators. The deliverable is also concerned with identifying use-case requirements for building the QUANTICOL tool integration platform.

Bus networks measurements are related to the City of Edinburgh. Data available to the general public offers model predictions of arrival times at bus stops. In addition, confidential data has been obtained from Lothian Buses, an Edinburgh-based company which operates a large bus network. It offers higher-quality information, consisting of detailed records of GPS locations, bus speeds, timestamps, and bus identifiers. Consistently with the proposal in Task 5.1a of the project's work plan, our preliminary analysis suggests that the quantity and quality of the data analysed thus far is sufficient to calibrate and validate certain spatial models of bus networks. In particular, this study has stimulated the development of patch-based models of bus routes which will further exercise the robustness of the data for model-building purposes.

For bike-sharing systems we focussed on publicly available data. Most systems offer web access to live information on the availability of bikes at parking stations. These websites can be systematically queried to reconstruct real availability traces at any desired granularity. Some operators have also provided historical traces consisting of journey details, including start/end dates and source/destination stations. Using the publicly available datasets for the bike sharing system of the City of London as a prototypical example, our data analysis confirms that these measurements can be used for spatial models with an explicit representation of the network topology.

Models of smart grids can benefit from the availability of measurements and forecasts at the transmission level, offering aggregated information about energy production, consumption, and market prices at the national electricity-network scale. Load measurements to calibrate models at a smaller scale (i.e., neighbourhood or building level) are also available in the research community. These are also accompanied by network benchmarks consisting of prototypical layouts that have been extensively studied in the literature. Overall, the available data is sufficient for the QUANTICOL purposes as discussed in Task 5.2a and demonstrated by a number of related papers already published. In the longer term, it will be possible to build large-scale models of adaptivity considering the short-term weather forecasts to drive adaptive control policies.

Finally, regarding tool development, an initial requirements-elicitation activity has highlighted a predominance of Java as the platform of choice for tool development by the project partners. This confirms the work plan set out in Task 5.3, affirming the use of a Java-centric QUANTICOL tool integration platform. This choice, however can still accommodate non-Java contributions, due to specific tool requirements or to the partners' expertise. Such an integration is possible, for instance, by exploiting a number of already available language bridging mechanisms such as Java Native Interfaces or Matlab's javabuilder.

In conclusion, the data at our disposal for our case studies are of appropriate quality for continuing our research along the lines of the proposed QUANTICOL work plan. Software integration will be facilitated by the emergence of a consensus around Java as the platform of reference for tool development.

Contents

1 Introduction

This deliverable discusses the progress made in the project regarding the analysis and the computer-assisted solution of models of the QUANTICOL case studies. In Year 1, the focus has been on problem-scoping and requirement-elicitation activities that will constitute the basis for further work in later stages of the project. With regard to model analysis, the objective of this report is to provide a discussion of data and measurements of the real systems which have been taken as case studies. In general, this study serves two distinct purposes in two different stages of a typical modelling workflow. It is essential to identify which modelling scenarios have the potential of being answered, since this is subject to the availability of data that can support the theory. For model validation, instead, data analysis is important to increase the quality of the collected measurements, the robustness of parameter-fitting techniques, and ultimately the overall predictive accuracy of the models.

We report on data analyses carried out for all our three case studies: smart buses, bike sharing networks, and smart grids. For each of them, after an overview of the system under consideration, we discuss requirements for modelling. In the case of smart buses these have been inspired by direct communication with service providers, namely representatives of the City of Edinburgh council and the Lothian Buses company, based in Scotland. They have provided a number of modelling questions as well as access to information and measurements that are not in the public domain, but which are necessary to parameterise and validate quantitative models of bus networks. In the case of smart grids and bike-sharing networks, instead, we consider modelling requirements that are shared by other researchers working in these fields. The long-term objective is to consider a common ground in order to understand how the novel techniques developed in QUANTICOL relate to the increasingly large research literature available.

For bus networks, we discuss two data sources. The first source is not publicly available and consists of timestamped records of GPS locations, from which it is possible to infer bus routes and durations of route segments. The second source is partially available to the public and consists of the real-time tracking data that is typically displayed at bus stops indicating the expected arrival time of buses. For smart grids, we consider a range of publicly available datasets which we characterise according to the scale of the measured system. We present aggregated data at the transmission level (i.e., national-scale electricity network) as well as load data at the local level (i.e., neighbourhood, building). We also report on ongoing work at EPFL to deploy a metering infrastructure for the EPFL grid releasing high-resolution load measurements into the public domain. For bike-sharing networks, we consider two datasets that are publicly available for the system operated by Transport for London. One dataset contains records of real-time availability information of bikes at every docking station collected through regular queries to the system's online interactive map. The other dataset consists of the details of every journey made between July 2012 and February 2013, allowing us to infer routes across the network and journey durations. For each case study, our data analysis is concluded with a discussion of related work, putting emphasis on which sources of data have been used in the literature, and for which reasons.

A complementary aspect reported in this deliverable is a requirement analysis for realising a QUANTICOL software framework for model specification and analysis. This has been achieved by aggregating the requirements of all project partners in terms of software tools, programming languages and libraries that have been or are envisaged to be used in QUANTICOL. This has led to the identification of a number of commonalities and shared concerns that, given the current requirements, allow us to consider the design of a Java-based framework.

The remainder of this report is structured as follows. Section 2 reports on data analysis on bus networks. Section 3 presents the data analysis of bike-sharing systems, while Section 4 discusses smart-grid data. Section 5 presents the results of requirements-elicitation activities for tool delevopment, while Section 6 concludes.

2 Urban Smart Transport Case Study

In this section we discuss the structure and visualisation of the various data sources used to build and parameterise our models for the urban smart transport case study. Our primary source of data here comes from our cooperation with the City of Edinburgh council and the Lothian Buses company, based in Scotland and operating an extensive bus network in Edinburgh.

This section is organised as follows. In Section 2.1 we provide a brief description of the stakeholders of the system and their goals and major indices of performance. Section 2.2 illustrates the requirements for this case study. In Section 2.4 we discuss the GPS data provided to us by Lothian Buses. In Section 2.5 we discuss the data that we obtain via the City of Edinburgh council's real-time prediction model, which consists of static topology information (e.g., bus routes and bus stop locations) as well as dynamic journey planning (predicting arrival times of buses at bus stops and journey times to later stops on the route).

2.1 Stakeholders

In an urban transport system there are many different types of stakeholders. There are the passengers who use the service, the operators who plan and adapt the service, the drivers who execute the service, and the regulators who stand outside the system and judge the service according to metrics, standards, and key performance indicators. Each of these has different criteria to evaluate the success of the service. As an example, we can identify the following points of view and success criteria:

- *Passenger-centric:* have as many buses as possible and as many routes as possible, have the waiting time at a stop approach zero, have the most efficient journey possible, have bus occupancy as low as possible (to increase the likelihood of finding a seat).
- *Operator-centric:* have as few buses as possible and as few routes as possible, have bus occupancy approach 100% (to increase the revenue earned versus fuel expended), have the most efficient service possible, have the most adaptable service possible.
- *Driver-centric:* have routes be as simple as possible, have incidents of overtaking approach zero, have typical bus occupancy well below 100% (to reduce the likelihood of having to turn passengers away because the bus is full), have the most predictable service possible.
- *Regulator-centric:* have as many buses as possible (to reduce the number of cars on the road), have the most efficient routes possible, have services run according to the published timetable, have carbon emissions approach zero, have noise levels approach zero.

Operating a successful urban transport system depends upon finding an acceptable compromise between the valid (but contrasting) points of view held by different stakeholders.

2.2 Requirements and Benefits of Modelling

We have undertaken the modelling of parts of the bus network located in Edinburgh and operated by Lothian Buses. Primarily we wish to model this system in order to evaluate proposed changes to the network without the cost of introducing the changes. Our models could ultimately be used to reason about which changes to the network are most beneficial to all parties.

- For example, there is an external proposal to reduce the maximum speed on roads within the city centre to 20 mph, a reduction of one third from the current 30 mph. Such a change may have a drastic effect on how well buses within the network can adhere to the published timetable. Our models could be used to support or oppose such a proposed change to the speed limit.

- Furthermore, bus networks are highly regulated with potentially severe (financial) consequences for an unacceptable level of performance, mostly with respect to when buses depart from their stops and how well that corresponds to the published timetable. Our models could be used to assist in determining what are fair performance attributes upon which to assess the bus company. They could also be used to assess whether or not a particular instance of failing to meet the requirements was within the control of the bus company.
- Changes to the network which respond to changing patterns in bus usage are a typical example of collective adaptation but it would be helpful to explore alternative scenarios before choosing the preferred modification. Our models could be helpful here.

2.3 Sources of Data

Four sources of data are available to us, in different quantities and different levels of detail.

- *AVL data from Lothian Buses* — Automatic Vehicle Location (AVL) data consisting of bus identifiers, a GPS location, and a timestamp.
- *Wi-Fi data from Lothian Buses* — AVL data consisting of bus identifiers, a GPS location, speed, heading, and a timestamp.
- *API data from "BusTracker"* — Available to developers with an Application Programmer Interface (API) key, subject to conditions of fair use. This data consists of model predictions of bus arrival times with bus identifiers included.
- *Open data from "BusTracker"* — Available to the general public, distributed over the World-Wide Web. This data consists of model predictions of arrival times. Bus identifiers are not included.

Below, we compare the relative availability of each source of data and the data content. Wi-Fi data is very informative, but it is not currently available in large volumes. Open data is available in essentially unlimited volumes, but it is not very detailed.

	Frequency	Resolution	Detail	GPS	Data	Volume
Wi-Fi	High (5 sec)	Lowest (min)	Highest	Yes	Raw	Lowest
AVL	Low (30 sec)	Highest (sec)	High	Yes	Raw	Low
API	Low (30 sec)	Low (min)	Low	No	Model	High
Open	Low (30 sec)	Low (min)	Lowest	No	Model	Highest

Despite differences in their availability and their scope, each of these types of data can be useful to the QUANTICOL project, either separately or in combination, for uses such as calibrating and parameterising models, dimensioning model scenarios, and identifying issues and problems where modelling can help to inform decision making. We consider them in more detail now.

2.4 Automatic Vehicle Location Data

The data provided to us by Lothian Buses consists of AVL data obtained using periodic GPS location measurements. Each data entry consists of a bus identifier, a location measurement and a measurement timestamp. We currently have access to three distinct datasets, namely:

1. a minor dataset consisting of location measurements of 15 buses of Route 10, collected between 31 May 2013, 12:07:32, and 6 June 2013, 17:04:23;
2. a major dataset consisting of 154 buses of various routes, collected between 9 August 2013, 15:31:58, and 10 August 2013, 18:49:13; and

3. a full fleet dataset which consists of 745 buses, collected between 28 January 2014 at 11:31:14 to 30 January 2014 at 12:38:31.

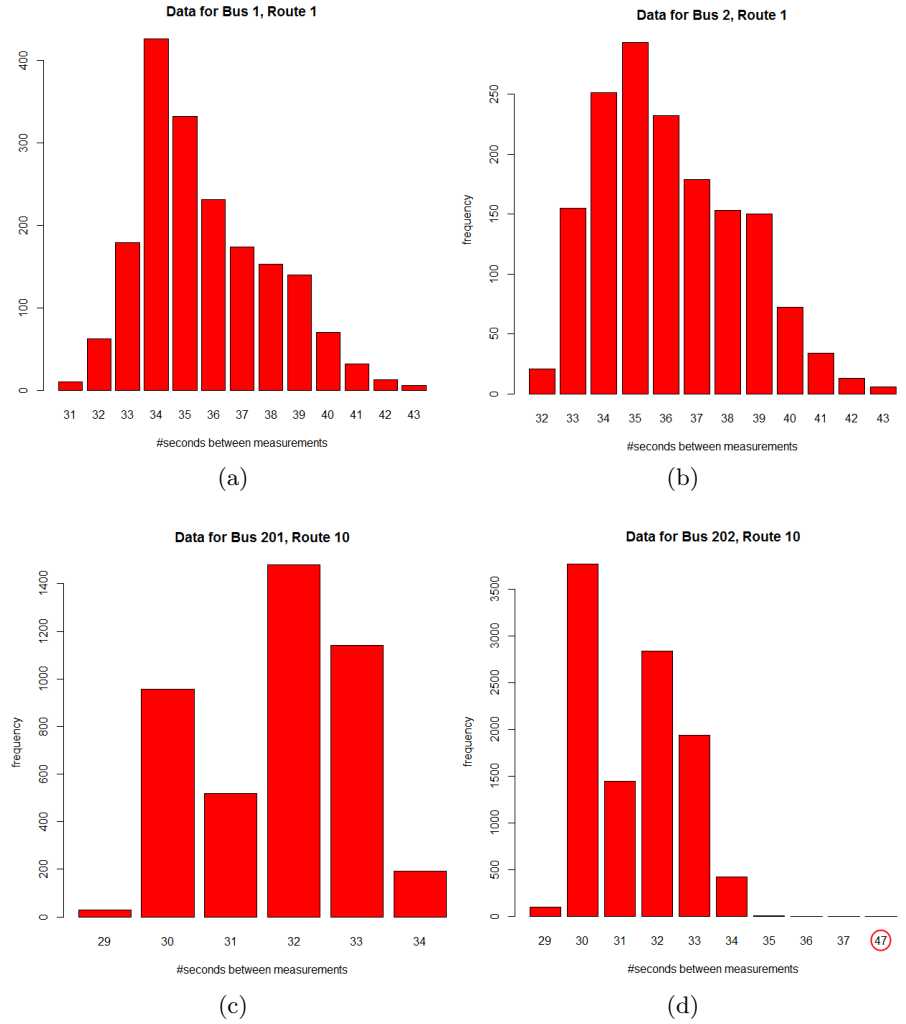


Figure 1: Distributions of the times between measurements for four different buses. Some distributions are heavy-tailed, for example in (d), a 47-second delay is observed: note the discontinuous axis.

Within a dataset, the measurements generally align, i.e., if a measurement exists for a certain point in time for one bus then a measurement should exist at the same point in time for all other buses in the same dataset. This was confirmed to us by Bill Johnston of Lothian Buses; data is gathered through centralised pull requests. The main exception to this rule is that when a bus has not moved between two measurements, the second measurement is not recorded (and similarly for larger sequences). The time between measurements is roughly between 30 and 40 seconds. Distribution plots are given in Figure 1. There does not seem to be an obvious correlation between the time between measurements and the time of day, although this has not been investigated further.

The data is not given in latitude/longitude coordinates but in Ordnance Survey for Great Britain (OSGB) eastings and northings. Once converted to latitude/longitude coordinates, the measurements have a horizontal (longitude) offset of about 0.00143 degrees, but this offset seems to be consistent across the different datasets. The GPS measurements are quite accurate once the offset is removed; we can determine the direction of the bus by the side of the road on which it is driving, illustrated in Figure 2.

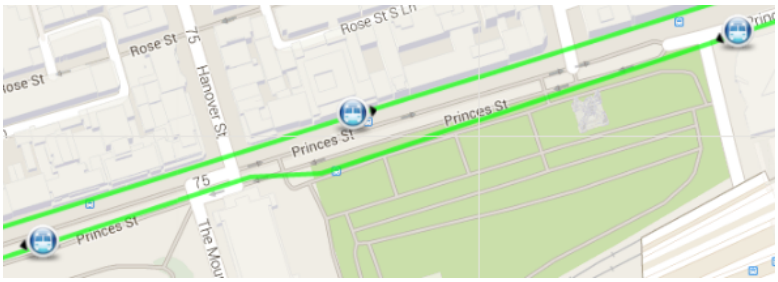


Figure 2: The data can be accurate enough to confirm the direction of the bus by inspecting visually the side of the road on which it is driving.

Some errors exist within the data: in extreme cases, this means that the location data is obviously invalid (e.g., buses having a location that corresponds to a patch of farmland or the middle of the Firth of Forth). In less obvious cases, buses seem to freeze for extended time periods before suddenly appearing in distant locations. Buses appearing to freeze in locations in the middle of the city may have a big impact on, e.g., running time analysis. Hence, it is advisable to have a human check the data for obvious errors using a visualisation tool.

While the data entries include a bus identifier, it is non-trivial to match a bus identifier to a bus route as the assignment of buses to routes may vary between days. Hence, the route of a bus must be inferred from the data. The bus routes can be obtained in term of sequences of coordinates using the MyBusTracker API.

In addition to the AVL data, Lothian Buses gathers location data using the WiFi system with which some of their buses have been equipped. An example is shown in Table 1. The location data gathered this way is not expressed in terms of eastings and northings but in latitude/longitude, so there is no for conversion. The WiFi data has a higher frequency than the AVL data; there are about 12 measurements per minute. A disadvantage of using this data is that the timestamps do not include seconds; if there are 12 measurements in the same minute, then all of these data entries have the same timestamp. If the location entries for two subsequent measurements is sufficiently different, then the speed and direction of the bus are included in the second entry. The speed data is expressed in miles per hour. In our sample dataset this varies between zero and thirty-seven miles per hour. Like the AVL data, the data is not entirely clean; there are sometimes large gaps between measurements and not all entries have values for all of the fields.

WAN Time	**Latitude**	**Longitude**	**Speed**	**Heading**
18/12/2013 18:32	55.961616	−3.187199	4.81636	75.1505
18/12/2013 18:32	55.961678	−3.1876917	5.65004	99.1059
18/12/2013 18:32	55.961107	−3.187367	6.52043	249.404
18/12/2013 18:32	55.961251	−3.187107	9.94455	224.96

Table 1: A sample of raw wi-fi location data for Edinburgh buses.

An advantage of AVL data is that it is amenable to *interpolation*. This can help to ameliorate the problems of low frequency data by allowing us to insert additional points between reported data. Using linear interpolation, we can insert a mid-point between adjacent points to generate (approximate) 18-second resolution data from (genuine) 36-second resolution data. Repeating this process, we can generate (approximate) 9-second resolution data from (approximate) 18-second resolution data. This interpolation is reasonable because buses are unlikely to be subject to rapid variations in their dynamics. Thus, we will supplement our reported data in this way where necessary in the belief that we will not be introducing significantly erroneous values through our use of linear interpolation.

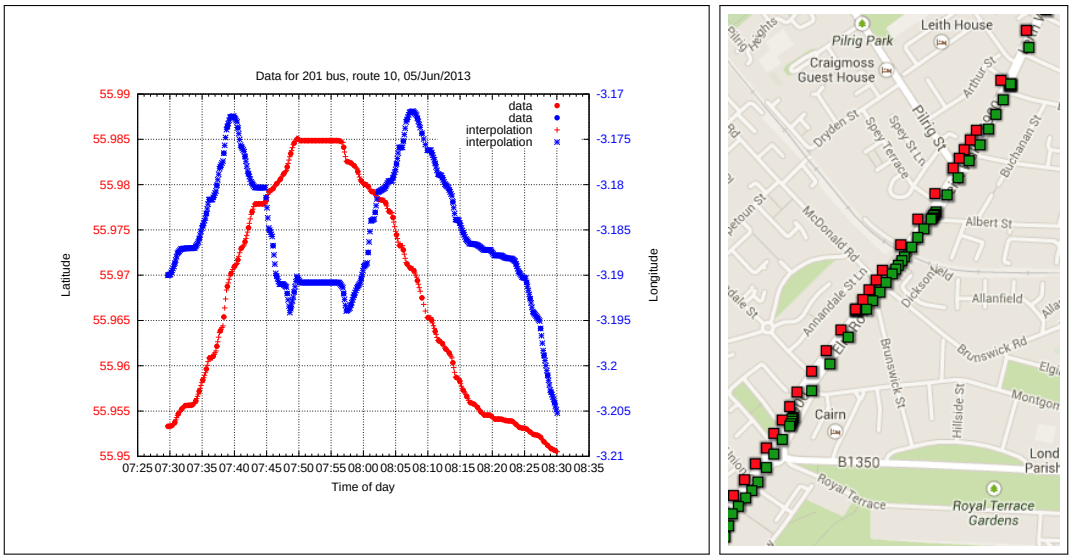


Figure 3: The same interpolated AVL data set is interpreted semantically on a graph as latitude and longitude time series (left) and rendered visually on a map (right). In the map visualisation the data points change from red to green at the mid-way point of the journey, representing the passage of time in an abstract sense, and explaining the symmetry in the graph by showing that the bus is returning along the route travelled in the outgoing part of the journey.

2.5 Real-time Tracking Model Data

The first stage in modelling the bus network has consisted of collecting data concerning bus journeys. We have focussed on two particular sequences of stops, one run by the 31 route and another by the 67 route. In cooperation with the City of Edinburgh Council, Lothian Buses provides a service, called *MyBusTracker*, for estimating when buses will arrive at each stop. MyBusTracker is available from a dedicated website[1] and presents the output of a prediction model which is fed with the AVL data on GPS positions of buses. This source of data has the benefit of being open and plentiful but it has the disadvantage that the data delivered in this way consists of model predictions rather than being raw measurement data of the kind which is available from the AVL system.

However, this information can be used to give us some insight into the stochastic fluctuations in the system caused by delays due to road congestion and unpredictable patterns of use which are dependent on passenger demand. For each stop, each service which uses that particular stop has an estimate for the minutes until the next bus arrives. When that estimate is lower than 3 minutes the time returned is simply rendered as the word "DUE". We can use this information to approximate when a particular bus has departed from a particular stop, since at that point the estimate for the next bus of the same service will jump from reading "DUE" to a number representing the time to wait for the next bus. In this way we can estimate the waiting time experienced by passengers. Figure 4 (left) shows the difference between this estimate and the expected waiting time published in the timetable for three selected bus stops, around midday on November 19th, 2013. It is possible to accumulate waiting time predictions for a series of stops along a route as in Figure 4 (right).

Initially this data was obtained by parsing the HTML code of the publicly accessible web page. The information content of this page is designed for use by passengers, which meant that we could only determine when *some* bus departed from a particular stop, not *which* bus it was. This meant that there was a manual stage in gathering together the departures from each stop undertaken by a single bus along the two sequences of stops on the two routes in which we are interested.

[1] `http://www.mybustracker.co.uk`

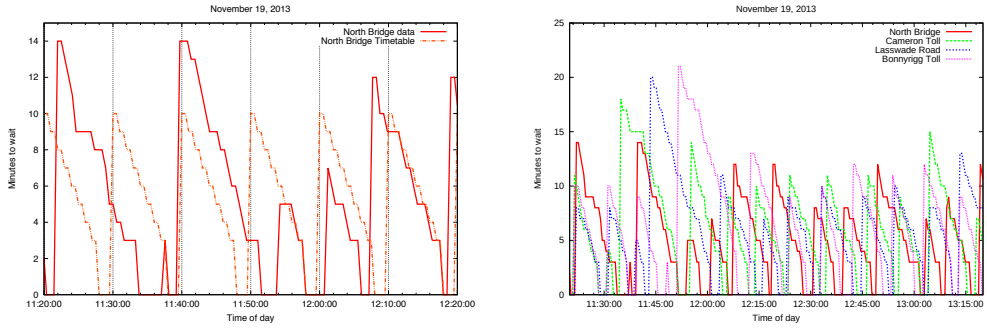


Figure 4: Left: Comparison of predicted passenger waiting times against the waiting time expected from the timetable. Right: Plot of predicted passenger waiting times along the 31 bus route.

We approached the City of Edinburgh Council to ask for greater access to the bus tracking data and have since received a developer key which allows us access to the non-public MyBusTracker API. The API can be used both to obtain live information about the expected arrival times of buses and static information about bus routes and bus stop locations. The functions in the API provide us the information of exactly which bus is expected at which stop and we can use this to approximate when each particular bus has departed from each particular stop and thus use this to reconstruct an entire journey made by a single bus.

2.6 Related Work

In this section we give a short overview of the state of the art in the field of using AVL data for the analysis of public transport performance. We mention a number of important publications and briefly summarise their contents.

The authors of [ZRW07] investigate the problem of deriving OD (Origin-Destination) matrices for public transport hotspots, i.e., matrices in which entry (i, j) corresponds to the number of passengers from stop i to stop j in a given time period. This may be of interest for the public transportation case study if passenger data becomes available. The authors use AVL and fare collection data of both trains and buses. One of the problems addressed in the paper is the fact that their data only covers check-*in*s. A methodology is developed to infer the alighting locations for 71.2% of the check-ins. Another challenge addressed in the paper is that the fare collection data does not include the bus stops at which the transaction occurs, but only the time of the transaction and a bus vehicle identifier. This is remedied using AVL data. Furthermore, bus stop locations are matched to railway stations, which is not always trivial.

The authors of [WAW11] apply the methodology of [ZRW07] to bus transit data in London. Their main data source consists of Oyster Card transactions, supplied to the authors by TfL (Transport for London). Using the transaction-matching methodology of [ZRW07], roughly two-thirds of check-ins can be matched to an alighting. The resulting OD-matrices (for several routes) are compared to survey data, giving a relatively good fit (no more than 4% off at each bus stop). The exchange of Oyster card data between TfL and MIT has also led to [Gor12], in which the methodology of [WAW11] is extended to include modes of transport other than buses, such as rail.

The authors of [CWCG11] study the problem of adjusting bus stop timetables such that the percentage of 'on-time' bus arrivals is maximised, assuming that changes in the timetables do not affect the behaviour of the buses. In the paper, a bus is assumed to be 'on-time' if its arrival occurs between X and Y minutes after the scheduled time. The authors assume that differences between scheduled and actual arrival of a bus to a stop are normally distributed, and based on this (or at least the normal distribution's symmetry) argue that it is optimal to choose the scheduled time of arrival

such that the expected time of arrival is $(X + Y)/2$ after the scheduled time. When the on-time interval is $(-2, 5)$, this basically means that buses are scheduled to be 1.5 minutes late on average. Empirical bus stop arrival times are obtained using AVL. The paper contains a case study with data obtained from Miami-Dade Transit (MDT), the main public transport operator in the area around Miami, Florida.

The authors of [YGW+13] study the problem of using low-frequency (in their case, once each 60 seconds) AVL data to estimate the difference between the times of two consequent arrivals of a bus of a certain route at a bus stop, called the 'headway'. To determine when a bus is at a stop, the 2-dimensional GPS locations are first projected onto a one-dimensional axis (this can be done without problems for 97.6% of all AVL data entries). Because the locations of the bus are only known at snapshots, the locations of the bus between the snapshots are interpolated. Instead of linear interpolation, the authors propose a method that includes a calibration phase (with a corresponding calibration dataset). In the calibration phase, the route is divided into 10-metre zones and the relative proportion of time in the dataset's time period that is spent in each of the 10-metre zones is calculated. The interpolation of all future bus movements is then chosen to match this distribution. After determining the bus arrival times using the interpolation, the empirical distribution of the headway at several bus stops is displayed and fitted to a 3-parameter gamma distribution, which is a well-known probability distribution from traffic analysis (this contradicts the normality assumption of [CWCG11]). The authors measure the goodness-of-fit of their gamma distribution using the Kolmogorov-Smirnov test statistic. A case study is performed using data for a single bus route in Boston.

Another noteworthy contribution is [FHMS06], which is a comprehensive technical report detailing the state of the art regarding the use of data from AVL systems or Automatic Passenger Counters (APC) to improve transit performance in 2006. Subjects discussed in the report range from a survey of the potential uses of AVL data to practical considerations. A number of case studies are discussed, namely Seattle, Portland, Chicago, New Jersey, and Minneapolis in the USA, Ottawa and Montreal in Canada and The Hague and Eindhoven in the Netherlands. The uses for AVL-APC data discussed by the authors include: analysis of running time, schedule adherence and headway regularity, demand analysis and bus route mapping.

The authors of [SF04] propose a model, based on a Kalman filter, that predicts bus arrival and departure times at stops. Particularly, their model separates bus running times from dwell times, and incorporates their interplay. Such interplay typically leads to alternating late and early buses: if a bus is late, it has to pick up passengers intended for the next bus, which means that the late bus gets further behind schedule and the bus after starts to run early because it has to pick up fewer passengers. This phenomenon is also mentioned in [BCBM07]. In [SF04], a case study is performed involving data provided by the Toronto Transit Commission.

2.7 Using Measurements for Modelling: Challenges

The main idea behind our formal model of bus movements in the city is to abstract the behaviour of the buses by dividing the bus routes into patches. Our model then describes the movement of buses between patches. An example of a division of a bus route into patches is displayed in Figure 5.

The choice of patches is non-trivial; e.g., one has the choice between a regular or an irregular grid (as discussed in Deliverable 2.1). The patch division of Figure 5 is irregular, and is based on the location of bus stops and traffic lights along the route.

As we can see from the log-scale histogram in Figure 5, the presence of traffic lights and, particularly, of bus stops in a patch has a significant impact on the distribution of the time spent by buses in that patch.

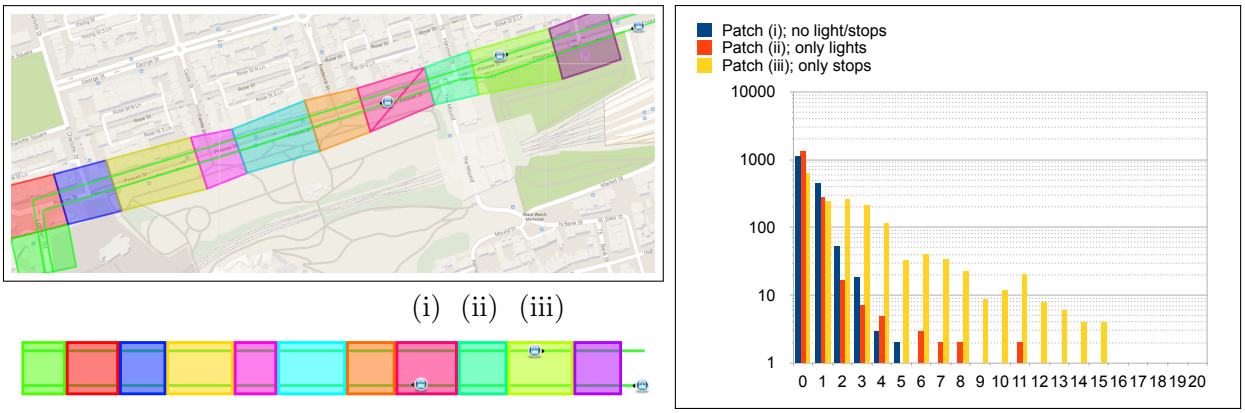


Figure 5: Left: Buses in two- and one-dimensional space; part of the route, including patches. Note that, in 1D, we assume for both directions that they cross the same amount of space in the leftmost red patch, although it is obvious from the 2D map that this is not the case. Right: Log-scale histograms of the number of roughly 36-second time slots spent in a patch: Patch (i) without traffic lights and without bus stops (blue); Patch (ii) *with* traffic lights and without bus stops (red); and Patch (iii) without traffic lights and *with* bus stops (yellow).

3 Bike Sharing Case Study

The goal of this section is to report on initial data-gathering activities regarding the case study on bike sharing, with the aim of identifying the sources of measurements that will be required to build and validate the models in later stages of the project. This contribution is organised as follows. Section 3.1 briefly overviews bike-sharing systems. Section 3.2 illustrates requirements and benefits for models of bike sharing. Section 3.3 discusses related work. Finally, Section 3.4 shows the results of data analysis, identifying a number of challenges that the measurements pose to model development and accuracy.

3.1 Description

Bike sharing systems (BSS) are becoming an increasingly widespread service of urban transport. They serve the main purpose of providing an environmentally friendly alternative to motorised transportation to reduce pollution levels in cities. Additionally, they offer a convenient complement to more traditional public services, in that they help tackle the so-called *last-mile problem*, concerned with effectively connecting public-transport passengers from bus/tram/train stations to their final destination (e.g., home, workplace, and so on). Although originally introduced as early as in the Sixties, *smart*, computer-assisted BSS have gained popularity since 2005, with the introduction of the Velo'v system in the city of Lyon [DeM09]. To date, there are over five hundred BSS worldwide, operating a roughly similar scheme that can be described as follows.

A BSS consists of a number of *stations*, i.e., parking places explicitly designed for bikes which are distributed across a city. Each station has a number of electronically controlled *docking points*, each accommodating a bike. An automatic machine allows the pick up of a bike from a docking point upon some proof of identity by the user (e.g., a smart card, a debit card, and so on). The user is charged according to the duration of the rental, which is measured at the time the bike is dropped off at a station. Knowing bikes' and users' identities discourages theft; furthermore, as will be discussed later in this document, it is possible for the provider to collect detailed usage statistics of the system. This information is essential for the model-building purposes of the QUANTICOL project. To improve users' experience providers usually make live information available on the web. In many cases this is achieved by providing an online interactive map which shows the location of the stations as well as their current availability and capacity. Figure 6, for instance, shows a screenshot of the online map

Figure 6: Screenshot from the live map at `https://web.barclayscyclehire.tfl.gov.uk/maps` providing live bike availability at London's Waterloo station.

of Barclays Cycle Hire scheme operated by Transport for London.[2] As will be discussed in the next section, availability traces can be obtained for statistical and modelling purposes.

As with the case study on the bus system of the city of Edinburgh, different stakeholders have somewhat conflicting expectations from a BSS. For example:

- *Users* would like to have as many stations as possible with as many docking points as possible. On the other hand, doing so could result in excessive costs borne by the *provider*. Thus, a trade-off must be found to ensure high service availability at an acceptable cost.
- Users that wish to *start* a journey would like to maximise the probability that a bike is found at the station where a bike is to be picked up. On the other hand, users that wish to *end* the journey at the same station would like to maximise the probability that an empty docking station is available.

3.2 Requirements and Benefits of Modelling

The QUANTICOL project will help set up these problems as rigorous mathematical problems. The main benefit is the possibility to examine what-if scenarios and carry out capacity planning without actually having to apply changes to the real BBS. In some cases, this is to be avoided because changes can be costly, e.g., if the provider would like to understand the impact of increasing the number of docking points at a station. In other cases, changes are even impractical to determine in practice; for instance, this is the case when one would like to anticipate the maximum tolerable population of users until the quality of service drops below some given threshold. Results of quantitative analysis might be useful not only for the service provider, but for users as well. For instance, a user wishing to use BBS as part of their commuting journey could be notified in advance of the most likely station to go to in order to satisfy their request. At the basis of all these investigations is the requirement of having available an accurate predictive model of a BBS. This objective can be broken down into two sub-tasks:

- Develop a model of a single BBS station *in isolation*. This represents a fundamental building block that abstracts away from topology considerations and is only concerned with identifying suitable abstractions to capture the contention for shared resources (i.e., bikes) by users. This initial requirement is likely to lead to models with a manageable number of parameters and of relatively low computational cost.

[2] Available at `https://web.barclayscyclehire.tfl.gov.uk/maps`.

- *Compose* models of isolated stations into network models, to take into account routing and to be able to study the repercussions that a change in a station has on other parts of the system. This requirement is likely to lead to models of very large scale for realistic cases, which will call for many of the abstraction techniques developed in this project.

In addition to representing a divide-and-conquer strategy to modelling and analysis, this is in line with the *compositional* flavour of the QUANTICOL approach, which looks at building complex models as a result of interactions between distinct components.

A proper validation of these problems will involve comparisons against real-world measurements. The remainder of this section describes the state of the art in the use of BSS measurements in the research literature, and an overview of the datasets that are envisaged to be used within QUANTICOL for validation.

3.3 State of the Art

BSS have become a popular topic of research. A number of works deal with availability data collected from regularly scraping the provider's online interactive map. The duration of the observation and the sampling frequency depend on the specific study under consideration and the procedure is implemented using tools that have not been made available to the public. Froehlich et al. study availability data of Barcelona's *Bicing* system, performing a cluster analysis and prediction using Bayesian networks [FNO09] using data obtained from the availability map.[3] Kaltenbrunner et al. [KMG+10] also deal with data from Bicing using the same approach, although measurements are used to train and validate a time series model. Lathia et al. study London's BSS to determine changes in usage patterns due to the introduction of a new access policy based on credit/debit-card identification, switched from an earlier policy based on an electronic proprietary key [LAC12].

In addition to availability traces, other works have also considered detailed per-journey datasets, typically containing information about start and destination stations, and start/end time of each journey. This data allows for different kinds of analysis because it presents topological, network-level information that is clearly not present in availability traces that consider stations in isolation. Because of the potential commercial value of this data, some operators do not make it publicly available, but may release it to researchers, as in the case of Borgnat et al. for Lyon's BBS [BAF+11], and Vogel and Mattfeld's study of Vienna's BBS [VM11]. Instead, Schuijbroek et al. use per-journey publicly available data from the BSS of Washington and Boston to study inventory rebalancing using a dedicated fleet of vehicles [SHvH13].

3.4 Data Analysis

In this section we present the results of a preliminary investigation of datasets of BSS. The purpose is to identify challenges that will need to be addressed in later stages of the project when developing predictive quantitative models. For this study, we considered the case of London's BBS. This was motivated by the availability of live data, as discussed above, as well as historical traces that are publicly available for download (upon registration) from the Transport for London's Developer's Area.[4]

Data sources. Historical traces contain detailed information about each journey made from 1st February 2012 to 2nd February 2013. The following information for each entry can be used to derive usage statistics and to validate models: i) trip duration (in seconds); ii) identifier of start station (integer number); iii) start date/time; iv) identifier of end station (integer number); v) end date/time. Additionally, the dataset contains the address, in textual form, of start/end stations. This information will not be used in the remainder of this section, however it can be exploited in later stages to develop

[3] Available at `https://www.bicing.cat/es/formmap`.

[4] Available at `http://www.tfl.gov.uk/businessandpartners/syndication/`.

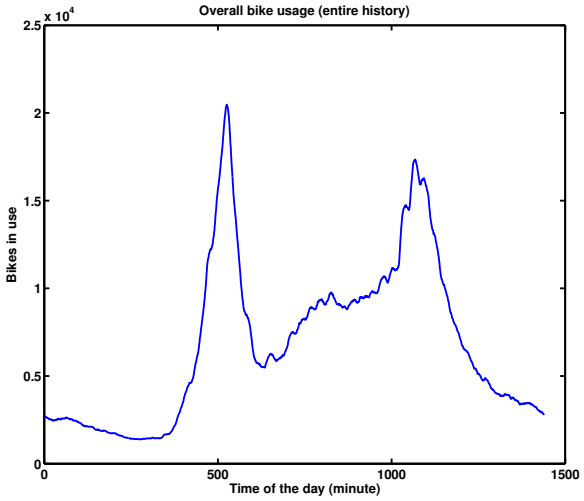


Figure 7: Cumulative bike usage between 1/5/2012 and 5/5/2012 as a function of the time of the day (expressed in minutes since midnight).

geographical models of the distribution of stations across the physical space. Live availability traces were collected by scraping the Transport for London web site. This was implemented with the help of a publicly available Java API.[5] In this section we report the analysis using data retrieved from 23rd October 2013 to 22nd November 2013. The website was queried at regular intervals every 20s. This sampling rate turned out to be sufficiently small to observe individual events of bike deposits or retrieval; less than 0.1% of the samples collected indicated changes in availability from the previous observation that were greater than one. For these, we arbitrarily assumed that individual events were spread evenly across the 20 second observation window.

Challenge 1: Time dependence. Using the historical traces, Figure 7 shows the cumulative usage of bikes between January 1st 2012 and January 5th 2012 as a function of the time of the day. In accordance with the literature (e.g., [LAC12]), the plot highlights a clear dependency of the system's behaviour on the time of the day, with relatively low usage at night and peaks during the rush hour (approximately at 9:00am and 5:00pm). Capturing this behaviour in an accurate manner represents a new challenge because most analytical results for quantitative models are available for the *time-homogeneous* case, where the dynamical laws are governed by coefficients (e.g., rates and probabilities) that *do not* depend on time.

Challenge 2: Heterogeneity. Figure 8 illustrates another important aspect of the BBS behaviour, namely heterogeneity both in *structural* and in *dynamical* properties of the stations. Figure 8a shows that stations have indeed different capacities, ranging between 10 and 64. Figure 8b shows the histogram of the mean interarrival times between successive events of bike drop-off, while Figure 8c provides the same histogram for bike pick-up events. In these two figures, only stations with at least 100 observations were considered, in order to increase the robustness of the estimated means. Figures 8a–8c have been obtained from the live availability traces. Using the historical traces, instead, Figure 8d shows differences in the routing behaviour. It plots the empirical cumulative distribution function of the average journey time between any two stations. This suggests that assumptions of homogeneity in the network routing may not turn out to be accurate when validated against the real system. Interestingly, we observe that a very large proportion of journeys (i.e., ca 93%) lasts less than 30 minutes. This can be explained by the fact that currently London's BBS offers a free allowance for

[5] Downloadable from from `http://bike-stats.co.uk`.

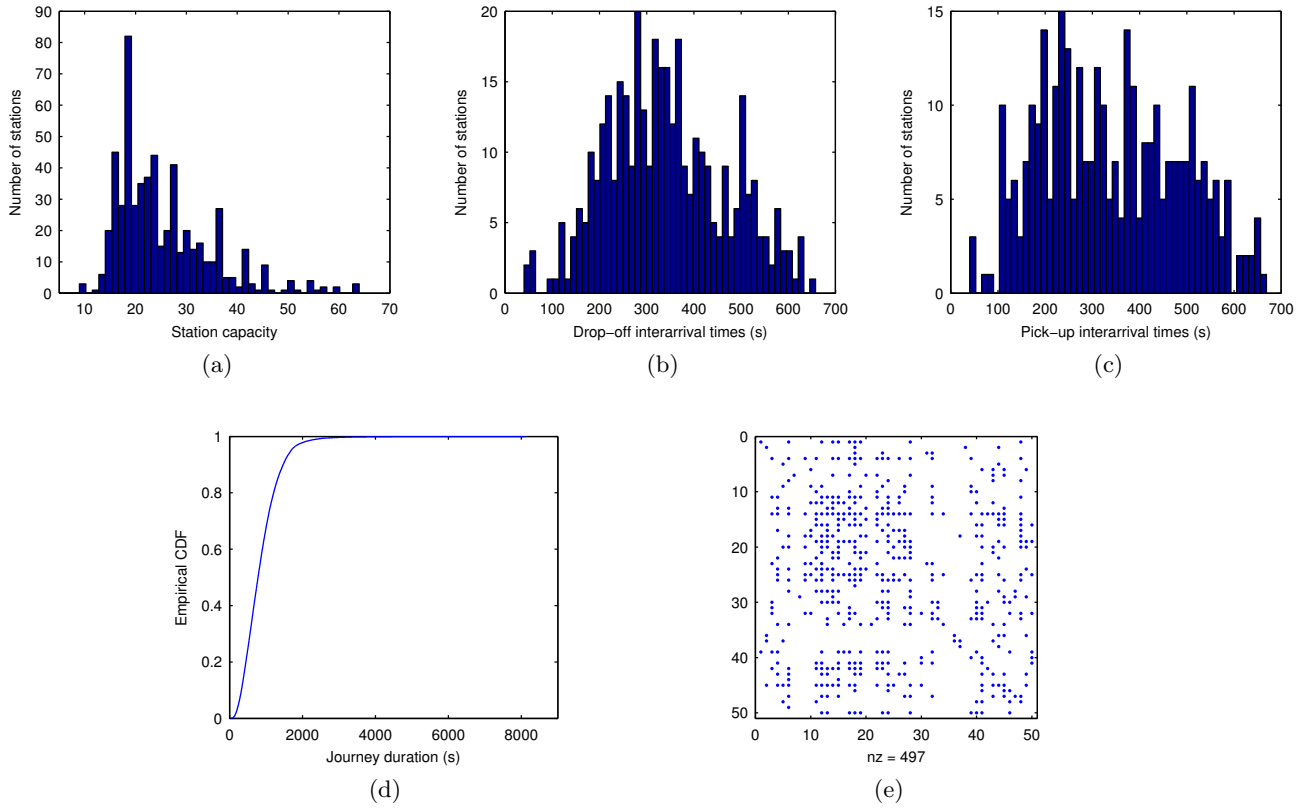


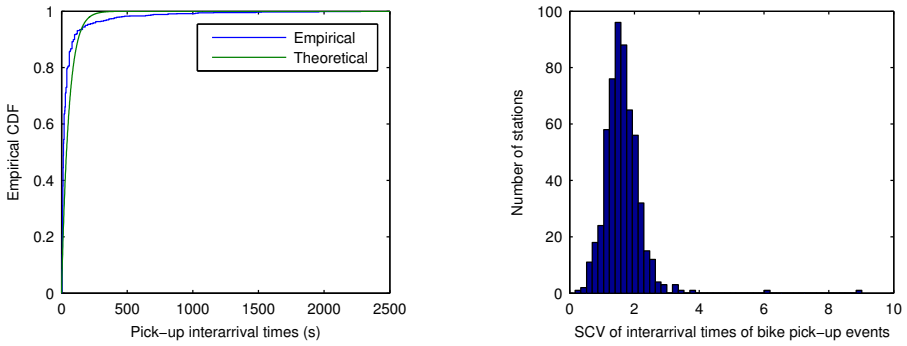
Figure 8: Heterogeneity in station behaviour.

bikes returned within that amount of time since pickup. Figure 8e is a graphical representation of the network's routing probability matrix for stations with identifiers between 1 and 50. A point is plot at coordinates (x, y) if there is a nonzero probability of going from station x to station y. This plot suggests that the routing is nonuniform, with a density of ca 20% for the number of nonzero entries in the matrix.

Challenge 3: Nonexponential distributions. In Figure 8 only the mean values of certain events have been shown. It is well-known that this information is sufficient to completely characterise a probability distribution if it is exponential. However, we now present some numerical evidence showing that this is not the case in the datasets considered here. For instance, using the live availability traces Figure 9a shows the empirical cumulative distribution function (blue line) of the interarrival times of bike pick-up events at the station with the highest squared coefficient of variation (SCV), equal to 9.02. For comparison, the theoretical cumulative distribution function of an exponential distribution with the same mean (equal to 54.95 s) is also plotted (green line). In fact, a great proportion of interarrival times is not exponentially distributed. This is shown in Figure 9b, which plots the histogram of the SCVs of the bike pick-up events across all stations. Here, only 6% of the stations (35/568) were found with an SCV within 10% of that of an exponential distribution (equal to 1.00).

4 Smart Grid Case Study

Electrical networks are evolving. Traditional energy systems are centralised, using accurate load prediction mechanisms with minimal storage and few monitoring points. Instead, future energy systems will incorporate more distributed and stochastic generation and feature two-way communication systems that enable real-time close-loop control. Renewable energy sources, such as wind and solar power,



(a) (b)

Figure 9: Nonexponential distributions.

are highly volatile and difficult to predict. For instance, current forecast techniques for wind power 12 hours in advance have normalised mean absolute errors of the order of 20% [CCN+08].

This section reports on the various data sets that we use in our smart grid case studies. We first describe in Section 4.1 the requirements in terms of data and the overall objective of our work. In Section 4.2 we describe the aggregated data that we use, and in particular the data about wind forecasting. Then, we discuss distribution networks in Section 4.3. Lastly, in Section 4.4, we discuss a project of metering the EPFL electrical grid that is currently deployed at EPFL.

4.1 Requirements and Benefits of Modeling

Our work on smart grids is mainly focused on how to deal with the stochastic nature of renewable energy. We aim to build control algorithms that help to incorporate these sources of energy. The nature of the problem studied spans two organisational levels. At the transmission level (national scale), we wish to study how to use demand-response or storage to provide energy balance and frequency regulation. At the distribution level (neighbourhood scale), we aim to develop algorithms which deal with local voltage or power violations.

To build these algorithms, we need to model the behaviour of the various components of an electrical network. Hence, one of our requirements has been to gather data about *generation*, both renewable and conventional generators, *transportation and distribution* networks, and *consumption* – in terms of elastic and inelastic loads. Our main sources of data are publicly available data sets. One of the main reasons is that, similarly to the bike-sharing data discussed in Section 3.4, energy providers do not publish data because of its potential commercial value. Furthermore, it can be used to target attacks against their system [BBH+12] and it can be considered to breach consumers' privacy [EK10]. Another reason is that many networks are poorly monitored, which means that the data is not available even to providers.

The main objectives of our data analysis can be decomposed in the following sub-tasks:

- **Forecast models** — Renewable energy is volatile and hard to predict. We use traces of wind production and weather forecast to build stochastic models of forecast errors. This allows us to build more robust policies and evaluate our control policies on real data (see Section 4.2 and Deliverable 1.1).

- **Spatial component** — The behaviour of the electrical grid is affected by its spatial component. For instance, the production of renewable energy is spatially correlated. Also, voltage or power flows are affected by constraints on the lines. We use available benchmark tools and measurements on campus to understand these effects (see Section 4.3 and Section 4.4).

- **Consumption modelling** — We wish to understand the consumption patterns and the amount of flexibility provided by electrical loads such as fridges or air conditioners. To achieve this goal, we rely on existing models. EPFL is also deploying a large metering infrastructure on campus to measure and control the consumption and the power flows on the campus' distribution network (see Section 4.3.2, Section 4.4 and Deliverable 1.1).

4.2 Aggregated Data and Forecasts

In many countries, the transmission system operators or the statistics office make public large data sets that contain data aggregated at a regional or national level. This data can include aggregated values of production and consumption, but also information on real-time prices. Most of the websites provide scripting mechanisms to recover large parts of the data set in batches.

4.2.1 Wind Forecast Data

The ELEXON portal[6] is a useful source of information for all data concerning the UK. We use it in several of our papers [GTLB12, GTLB14a, GTLB14b]. This portal provides operational data about UK grids. This archive is composed of daily reports that contain imbalance volumes, historic system price and, and operational data. We specifically consider the values of aggregated electricity production and consumption in the UK, and actual production or forecast value of wind electricity. These values are averaged over 30min intervals. In [GTLB12, GTLB14b], we use wind production and day-ahead wind production forecast in the time interval from June 2009 to April 2012, to build a stationary model of wind forecast error. These data allowed us to compare the performance of various scheduling policies using three typical wind forecasts:

- Persistence forecast W^{NP}: the forecast for production is equal to the average value of wind power during the previous hour. This forecast is widely studied in the literature [HM11, BDNL08] and performs well on short time scales [CCN$^+$08].
- Day-ahead forecast W^{DA}: the forecast for a day is made the day before at 3pm. The data are included in the data archive.
- Corrected forecast W^{LC}: we have developed this new method that uses the day-ahead forecast as an input and apply a linear filter that uses the fact that forecast errors are statistically corrected.

To illustrate these forecasts, we plot a typical sample of the actual wind production and the two forecasts W^{NP} and W^{LC} on Figure 10. The forecast W^{LC} is often better than W^{WP}, but sometimes it over-corrects (*e.g.*, on March, 21st). We also show the normalizsd mean absolute error on Figure 11 as a function of the forecast horizon i. The performance of W^{WP} does not depend on the time horizon considered because it uses day-ahead forecasts. The persistence forecast outperforms weather predictions for short time horizons (i.e., predictions less than 6h in advance). Our own forecast method, W^{LC}, provides the best of the two worlds. On average, it always outperforms the two other methods. We also evaluated more complicated statistical corrections (non-linear or with larger time-horizons) than W^{LC}, but they were not found to be significantly better.

We have extended this methodology in [GTLB14b], to construct a non-stationary model of wind forecast errors. We use this data to numerically evaluate the effect of elastic demand on the prices of real-time electricity markets. Our method is based on Pinson's methodology [PMN$^+$09] and uses multiple trajectories to represent forecasts. The trajectories are generated using the empirical covariance of the forecast errors. This methodology creates a realistic and tractable trajectorial forecast error model where trajectories separate into branches over time.

[6] https://www.elexonportal.co.uk/

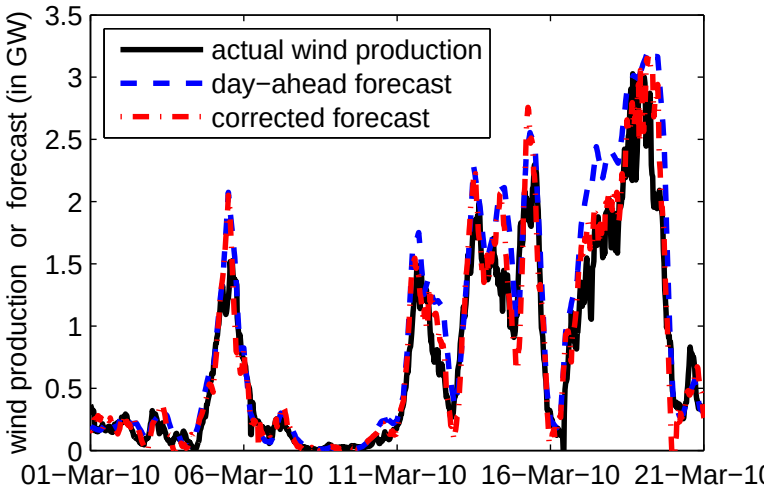
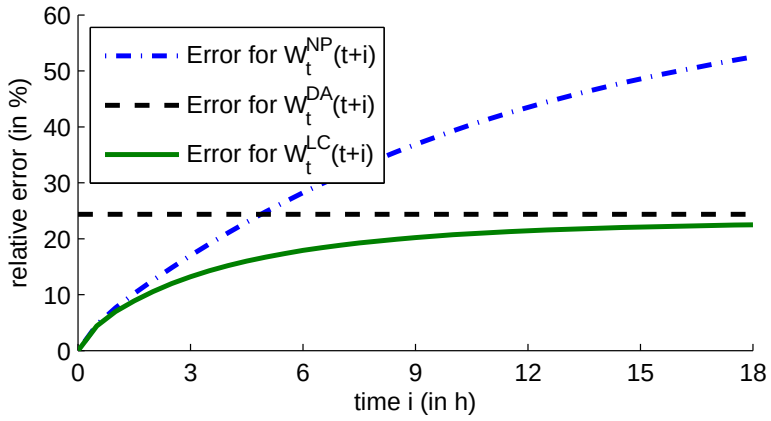


Figure 10: Wind forcast and production at the scale of UK. Typical sample of day-ahead forecast W^{WP} and forecast with linear correction W^{LC} versus actual wind power production (March 2010). Courtesy of [GTLB14b].



The x-axis represents the time-horizon i of the forecast. If $W^f(t+i)$ is the forecast done at time t for time $t+i$ and $W(t+i)$ the actual wind production at time $t+i$, then the mean absolute error is

$$\frac{\sum_t |W_t^f(t+i) - W_t(t+i)|}{\sum_t W(t)}$$

Figure 11: Mean absolute error of the three wind forecast definitions.

4.2.2 Other Statistics

A central question of our research is to quantify the amount of storage or flexible loads needed to provide a given level of service. Hence, to build realistic models it is necessary to use statistical data that can guarantee that desired targets are realistic.

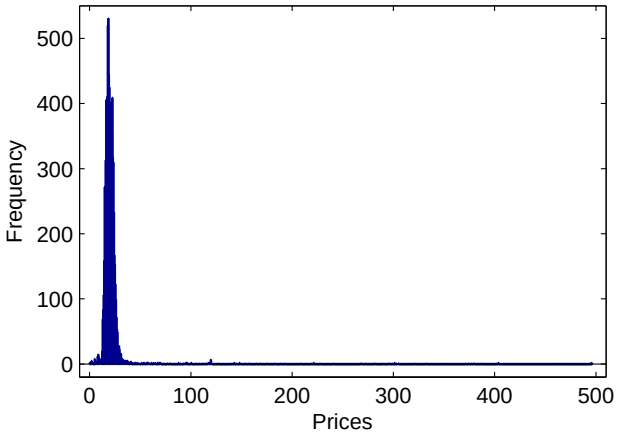
For example, the Swiss Federal Statistics Office provides information about energy generation and consumption in Switzerland.[7] The historical data are provided, as well as projections for up to 2050. Similarly, the DECC calculator[8] provides data about current wind production in the UK as well as different scenarios for consumption or generation of electricity in the UK until 2050. These websites also provide an aggregated view of how much storage is present in the network. It is also sometimes useful to zoom in on one particular generation unit. The big generation units often provide these information online.[9]

[7] `http://www.bfs.admin.ch/bfs/portal/en/index.html`

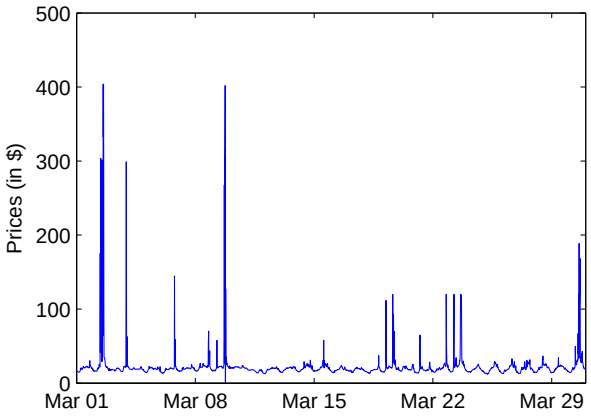
[8] `http://2050-calculator-tool.decc.gov.uk`

[9] For example in the US `https://www.dom.com/about/stations/hydro/bath-county-pumped-storage-station.jsp`, or in the UK `http://www.fhc.co.uk/dinorwig.htm`

For transparency reasons, operational data about markets and prices are also available online. These datasets usually contain historical and real-time prices, as well as generation and imbalance volumes. For example, in the US two extensive data sets that can be accessed are the Texas Market[10] and the PJM interconnection.[11] For example, we plot real-time prices that are observed on real-time Market in Texas for the beginning of 2012 in Figure 12. The distribution of prices is shown in Figure 12a. This graph shows that prices tend to concentrate around the average value. However, when we look at the evolution of prices over time (Figure 12b), we observe that they exhibit large fluctuations. These values are explained by theoretical models of real-time prices developed by us and others [WNPK+11, GLBPT13]. Similar data can be found for Europe, for the continental market[12], or for the Nordic Energy Market.[13]



(a) Distribution of real-time Market prices at the Houston hub during January-March 2012.



(b) Real-time prices at the Houston hub in Texas during the month of January 2012. Other months are similar.

Figure 12: Real-time market prices in Texas.

4.3 Network Modelling

To evaluate local scenarios, i.e., situations at the distribution level, we rely on two kinds of data: network distribution models and load traces. The networks considered here are representative scenarios that are available in the literature. The load traces are obtained from data available in related papers.

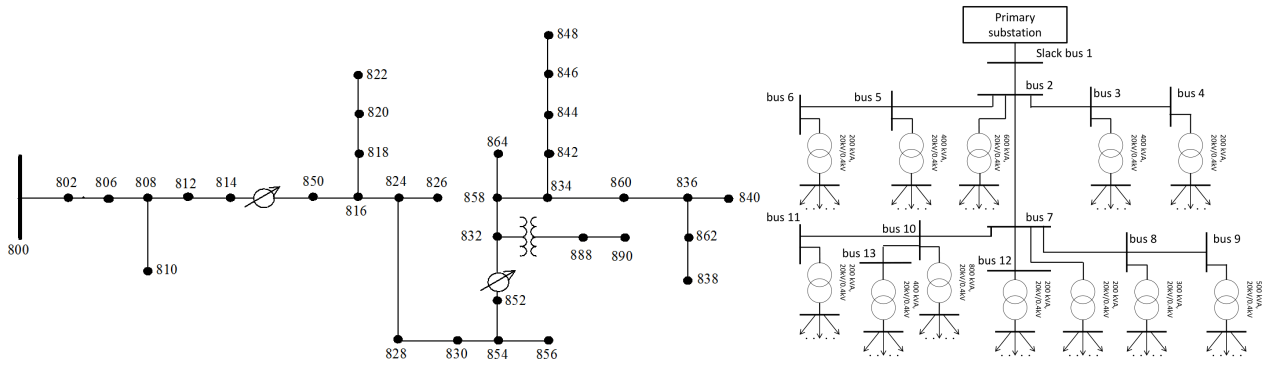
4.3.1 Network Distribution Models: The IEEE Test Feeders

Since 1991, the *IEEE distribution test feeder working group* has made available a number of network topologies. These topologies represent typical radial distribution network. Most (if not all) of the papers that evaluate the effect of consumption or generation on distribution networks use these topologies as test-cases. In the power systems community, it is mandatory to use these test-cases to evaluate a new control strategy. They serve as benchmark and avoid each paper using its own assumptions for line or load models, so that the results of various programs can be compared.

These distribution networks are described in [Ker01] and available online.[14] The description of each scenario includes the parameters the lines and typical load profiles. They can be used to evaluate voltage deviations on a node or power limits of lines. For example, [CTLBP12] explores a voltage

[10] `http://www.ercot.com/mktinfo/prices/`
[11] `http://www.pjm.com/markets-and-operations/etools/oasis.aspx`
[12] `http://www.transparency.eex.com/en/`
[13] `http://www.nordpoolspot.com/`
[14] `http://ewh.ieee.org/soc/pes/dsacom/testfeeders/`

control mechanism based on the broadcast of low-bit rate control signals. Examples of test-feeder are shown on Figure 13.



(a) The IEEE 34 test feeder (courtesy of [Ker01])

(b) The IEEE 13 test feeder, adapted to the load modeling of [CTLBP12].

Figure 13: Example of test feeders used to evaluate voltage.

4.3.2 Load Data and Elastic Loads

The IEEE test-feeders contain typical load profiles that can be used to model the basic consumption of houses. In our work, we aim to use models of elastic consumption that can be found in the literature. For example, the authors of [KMC11, MBBE13] develop a simple but realistic model of on/off loads. This model can represent thermostatical loads – e.g., fridges, AC, heating systems – or any load that has a duty cycle – e.g., cleaning a swimming pool. The common denominator of these loads is that their consumption can be anticipated or delayed, but not indefinitely – e.g., a fridge has a minimum and a maximum temperature. These models can be easily parameterised using measured data [KSAC11, KT12, Section V]. For instance, we used a variant of the elastic model of [KMC11] in [GTLB14a].

4.3.3 Generation: Solar Data

In order to study the effect of distributed generation on distribution networks, we used real traces of data of wind or solar electricity production. These data traces are the measured production of a single photovoltaic panel or a small installation. They cannot be obtained from national data, that contain only aggregated values. We mainly used two sources of data. The first source is [LKAC12], which provides one-second resolution data of power production. A part of the data set is accessible online.[15] The second source comes from weather wireless stations that are installed at EPFL, containing production of the photovoltaic panels installed on the roof of EPFL buildings.[16]

4.4 Deployment of a Metering Infrastructure of the EPFL Grid

EPFL is planning to deploy a smart grid infrastructure on campus. The goal is to equip all medium voltage transformers with Phasor Measurement Units (PMUs) and deploy a communication network infrastructure to centralise these data to a electrical network state estimator. Later, this state estimator will be used to experiment on demand-response and storage management policies. Figure 14 gives an idea of the communication network that will be deployed.

[15] `http://maeresearch.ucsd.edu/kleissl/demroes/`
[16] `http://lcav.epfl.ch/sensorscope-en`

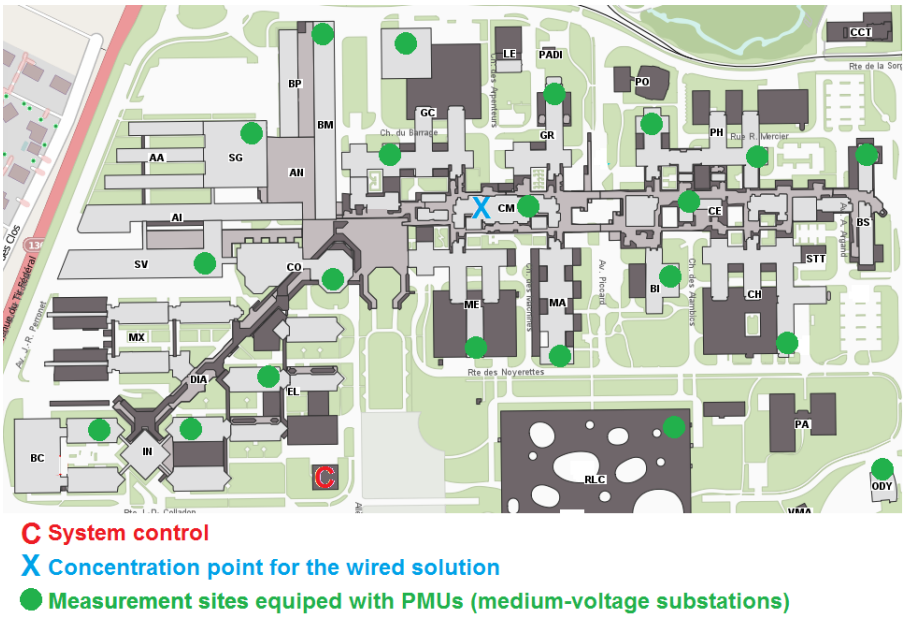


Figure 14: Map of EPFL and envisaged places of PMU measurement units that will be deployed. Courtesy of [PGTLB13].

The first step of this project aims at monitoring the state of the electrical network of EPFL in real time. The historical values of phases, voltage and state estimation will then be made available to the rest of the research community with a time resolution of 50ms. One of the main goals of this activity is to collect and make available to the public, for the first time, a dataset which will contain several years of data about a large and real grid with a very fine-grained resolution. This project will also serve as an experimental deployment of a real-time communication infrastructure for the smart-grid. In the beginning, existing twisted pair cabling will connect measurement and control endpoints. Later on, wireless infrastructure will be designed to facilitate system deployment on sites which, unlike EPFL, do not have widely deployed cabling.

5 Tools

This section discusses the use-case requirements for building the tool integration platform for QUANTICOL. These were obtained in two phases. First, we identified the set of software tools, libraries, and programming languages that have been or are intended to be used across all the work-packages. Then we considered which tools are envisaged to require interoperability. Based on this choice, we developed a preliminary plan to implement interoperability in a common framework.

Use cases were elicited by compiling a simple form containing the following information for each software tool that is supposed to be integrated.

Name This should contain any further information to uniquely identify the tool, such as version number in the case of an explicit dependency on a specific release.

Platform Operating system for which the tool is available.

Licence Software licence under which the tool is released. This information is needed to develop a licensing policy for the QUANTICOL framework.

Description Brief description of the tasks to run. This was used to identify possible alternatives and shared concerns across collaborators.

Name	*Platform*	*Licence*	*Description*
BioPEPA Eclipse Plugin	Eclipse Java	GNU	Implementation of the stochastic process algebra BioPEPA. Shares analysis library (stochastic simulation and fluid analysis) with the PEPA Eclipse plugin. `http://homepages.inf.ed.ac.uk/jeh/Bio-PEPA/biopepa.html/`
jSAM	Java	Eclipse	Integrates different specification languages and model-checking algorithms. `https://code.google.com/p/jsam/`
Matlab	Linux OS X Windows	Matlab	Rapid prototyping of models, linear algebra (numerical solution of Markov chains), numerical integration of ordinary differential equations (in particular, *stiff* solvers), statistics, parameter fitting, linear and nonlinear optimisation. `http://www.mathworks.com/`
MultiVeStA	Java	UIUC	A statistical analysis tool which can be easily integrated with discrete event simulators, enriching them with efficient distributed and statistical analysis and statistical model checking capabilities. `https://code.google.com/p/multivesta/`
OCaml	Linux OS X Windows	Q	Ease of translating logical and mathematical reasoning into fast, executable code. Useful to implement the semantics of small programming languages and logics (e.g. model checkers). `http://ocaml.org/`
PALOMA	Java Matlab	—	Prototype implementation of the process algebra PALOMA (see Deliverable 4.1), supporting discrete-event simulation and an ordinary differential equation solver which is based on Matlab.
PEPA Eclipse Plugin	Eclipse Java	GNU	Implementation of the stochastic process algebra PEPA as an Eclipse plug-in. Used for rapid development of models, stochastic simulation, and fluid analysis. `http://www.dcs.ed.ac.uk/pepa/tools/plugin/`
SimHyA	Java	GNU	A library which supports hybrid analysis, stochastic simulation, and fluid and linear noise approximations. Preliminary support for statistical model checking.

Table 2: List of tools for QUANTICOL.

The list of tools identified so far for the QUANTICOL project is shown in Table 2. Some of these have already been used in QUANTICOL publications. More specifically, Matlab has been used for stochastic simulation and the numerical solution of ordinary differential equations (ODEs) for queueing network models in [TT13]. In [BT13] an experimental numerical assessment compares the performance of stiff Lipschitz continuous ODE models against equivalent non-stiff ODEs with discontinuities. In [Tri14], Matlab's genetic algorithm implementation is used to study the effectiveness of an approach that prunes the parameter space of an optimisation problem for fluid models using results of monotonicity of differential equations. Matlab and the Bio-PEPA Eclipse plugin are used in [BLM13] to perform stability analysis of fluid models. MultiVeStA has been presented in [VS13].

To date, the issue of interoperability appears to be best tackled by envisaging a Java-centric QUANTICOL framework for Java. This is motivated by the ubiquity of Java in any platform, and by the fact that the majority of the listed tools are natively Java based. In addition, Matlab enjoys easy interoperability with Java thanks to the *Matlab Java Builder*,[17] a toolkit to export Matlab programs as Java classes. Interoperability with other software that does not directly communicate with Java, e.g., OCaml, can be achieved by creating ad hoc Java wrappers that invoke executables.

[17] `http://www.mathworks.co.uk/products/javabuilder/`

6 Conclusion

This deliverable has reported progress made within Work Package (WP) 5 during the first twelve months of the project. It has covered Tasks 5.1a and 5.2a, concerned with the case studies on smart public transportation and smart grids, respectively, and Task 5.3, about tool integration.

The intent during the first phase of Tasks 5.1a and 5.2a (which will be running until M24) was problem scoping and data gathering. In this deliverable, these activities have resulted in a preliminary analysis of the sources of data available and in the identification of a number of modelling scenarios to be investigated in later phases of the project. As a general remark, it is possible to observe that public availability of data is mainly hindered by the potentially commercial value that stakeholders attach to this information. However, we believe that in QUANTICOL a number of useful countermeasures have been taken in order to tackle these issues. The close collaboration with the City of Edinburgh Council and Lothian Buses has allowed us to obtain privileged information which is only partially in the public domain. This is in addition to a wealth of resources that are already available. The situation is somewhat more favourable for the bike sharing case study, where the publicly available datasets appear appropriate for capturing essential network dynamics. For the case study on smart grids, the availability of public data on aggregated energy production/consumption, benchmark network distribution models, and in-house smart grid deployments planned at EPFL can cover for a significant amount of data needs for model development, calibration and validation.

Together with progress made in other work packages, our preliminary analysis has stimulated interesting and challenging research questions. More specifically:

- The bus network datasets suggest the use of models with an explicit representation of space, where bus stations are represented as discrete nodes. In particular, this has benefitted from the taxonomical study conducted in WP 2 (and presented in Deliverable 2.1), which has helped us identify patch-based abstractions as an appropriate modelling technique to investigate. The availability of data on journey times allows us to build compelling quantitative models of bus routes, which can be used in simulation studies or model-checking studies to determine the whether timetabling or other issues could be improved.

- The bike-sharing datasets will be used to calibrate a network model of a bike-sharing system using queueing theoretic models with mean-field or moment-closure approximation (see Deliverable 1.1 for an overview of these techniques). Even using these approximations, a detailed representation of a real network topology such as London's bike sharing system leads to a large-scale model. This provides a case study for Task 3.2 in WP 3, where notions of aggregations beyond population models are being investigated. An application of these models is capacity planning, which involves varying the parameters of a model in order to find the best configuration that optimises a given cost function. In this respect, we will also take advantage of the work made in the context of Task 3.3, about efficient variability analysis.

- The smart-grid datasets offer the opportunity to develop models which integrate forecasts of volatile energy sources with adaptive control of energy demand. We will pursue the work started in WP 1 and use this data to design optimisation techniques for the problem of decision making in the presence of uncertain forecast. The different data sets covered in this deliverable will be useful to quantify the advantages of our policies in realistic scenarios. This data also shows us what are the requirements for the model-building activities in WP 1, in particular regarding uncertainties and multiple scales. The datasets also present a spacial correlations that need to be taken into account. This will drive the interactions between WP 2 and the smart-grid case study.

Deliverable 2.1 reports preliminary modelling work using our case studies with emphasis on spatial representation.

Work on Task 5.3 has regarded the requirements for tool development and integration. An initial elicitation has resulted in a number of different platforms and tools of choice, motivated by the variety of concerns and research expertise within QUANTICOL. However, building a Java-based platform seems to directly satisfy the majority of needs and can also accommodate integration with other platforms. New prototype implementations have been developed during the first twelve moths of the project: FlyFast, part of the jSAM tool suite and discussed in Deliverable 3.1, is a model-checker for mean-field models; PALOMA (discussed in Deliverable 4.1) is ongoing work on the development of stochastic process algebra with an explicit representation of space and is supported by a Java tool. As set out in the work plan, our overall objective is to integrate these contributions in tool chains. Toward this goal, future work has been planned on identifying opportunities for tool interaction. For example, work is under way to integrate the model-checking capabilities of MultiVeStA (discussed in Deliverable 3.1) with the Bio-PEPA Eclipse plug-in.

References

[BAF+11] P. Borgnat, P. Abry, P. Flandrin, C. Robardet, J.-B. Rouquier, and E. Fleury. Shared bicycles in a city: a signal processing and data analysis perspective. *Advances in Complex Systems*, 14(3):415–438, 2011.

[BBH+12] Andrey Bernstein, Daniel Bienstock, David Hay, Meric Uzunoglu, and Gil Zussman. Power grid vulnerability to geographically correlated failures-analysis and control implications. *arXiv preprint arXiv:1206.1099*, 2012.

[BCBM07] Mathew Berkow, John Chee, Robert L Bertini, and Christopher Monsere. Transit performance measurement and arterial travel time estimation using archived AVL data. *ITE District 6 Annual Meeting*, 2007.

[BDNL08] Hans Bludszuweit, José Antonio Domínguez-Navarro, and Andrés Llombart. Statistical analysis of wind power forecast error. *Power Systems, IEEE Transactions on*, 23(3):983–991, 2008.

[BLM13] Luca Bortolussi, Diego Latella, and Mieke Massink. Stochastic process algebra and stability analysis of collective systems. In Rocco De Nicola and Christine Julien, editors, *Coordination Models and Languages*, volume 7890 of *Lecture Notes In Computer Science*, page 1–15. Springer Berlin Heidelberg, Springer Berlin Heidelberg, 2013.

[BT13] Luca Bortolussi and Mirco Tribastone. Differential analysis of interacting automata with immediate actions. In *7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, 2013.

[CCN+08] Alexandre Costa, Antonio Crespo, Jorge Navarro, Gil Lizcano, Henrik Madsen, and Everaldo Feitosa. A review on the young history of the wind power short-term prediction. *Renewable and Sustainable Energy Reviews*, 12(6):1725–1744, 2008.

[CTLBP12] Konstantina Christakou, Dan-Cristian Tomozei, Jean-Yves Le Boudec, and Mario Paolone. Gecn: Primary voltage control for active distribution networks via real-time demand-response. *Smart Grid, IEEE Transactions on*, 99, 2012.

[CWCG11] Fabian Cevallos, Xiaobo Wang, Zhenmin Chen, and Albert Gan. Using AVL data to improve transit on-time performance. *Journal of Public Transportation*, 14(3):21–40, 2011.

[DeM09] Paul DeMaio. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation*, 12(4), 2009.

[EK10] Costas Efthymiou and Georgios Kalogridis. Smart grid privacy via anonymization of smart metering data. In *Smart Grid Communications (SmartGridComm), 2010 First IEEE International Conference on*, pages 238–243. IEEE, 2010.

[FHMS06] Peter G. Furth, Brendon Hemily, Theo H.J. Muller, and James G. Strathman. *Using archived AVL-APC data to improve transit performance and management*, volume 113. Transportation Research Board, 2006.

[FNO09] J. Froehlich, J. Neumann, and N. Oliver. Sensing and predicting the pulse of the city through shared bicycling. In *Proceedings of the 21st International Jont Conference on Artifical intelligence*, IJCAI'09, pages 1420–1426, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

[GLBPT13] Nicolas Gast, Jean-Yves Le Boudec, Alexandre Proutière, and Dan-Cristian Tomozei. Impact of storage on the efficiency and prices in real-time electricity markets. In *Proceedings of the fourth international conference on Future energy systems*, pages 15–26. ACM, 2013.

[Gor12] Jason B. Gordon. Intermodal passenger flows on london's public transport network. Master's thesis, MIT, 2012.

[GTLB12] Nicolas Gast, DanCristian Tomozei, and JeanYves Le Boudec. Optimal storage policies with wind forecast uncertainties. *ACM SIGMETRICS Performance Evaluation Review*, 40(3):28–32, 2012.

[GTLB14a] Nicolas Gast, Dan-Cristian Tomozei, and Jean-Yves Le Boudec. Impact of demand-response on the efficiency and prices in real-time electricity markets. *submitted*, 2014.

[GTLB14b] Nicolas Gast, Dan-Cristian Tomozei, and Jean-Yves Le Boudec. Optimal generation and storage scheduling in the presence of renewable forecast uncertainties. *accepted in Transaction on Smart Grid*, 2014.

[HM11] B Hodge and Michael Milligan. Wind power forecasting error distributions over multiple timescales. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–8. IEEE, 2011.

[Ker01] William H Kersting. Radial distribution test feeders. In *Power Engineering Society Winter Meeting, 2001. IEEE*, volume 2, pages 908–912. IEEE, 2001.

[KMC11] Stephan Koch, Johanna L Mathieu, and Duncan S Callaway. Modeling and control of aggregated heterogeneous thermostatically controlled loads for ancillary services. In *Proc. PSCC*, pages 1–7, 2011.

[KMG+10] A. Kaltenbrunner, R. Meza, J. Grivolla, J. Codina, and R. Banchs. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. *Pervasive Mob. Comput.*, 6(4):455–466, August 2010.

[KSAC11] Taylor M Keep, Froylan E Sifuentes, David M Auslander, and Duncan S Callaway. Using load switches to control aggregated electricity demand for load following and regulation. In *Power and Energy Society General Meeting, 2011 IEEE*, pages 1–7. IEEE, 2011.

[KT12] George Koutitas and Leandros Tassiulas. A delay based optimization scheme for peak load reduction in the smart grid. In *Proceedings of the 3rd International Conference on Future Energy Systems: Where Energy, Computing and Communication Meet*, page 7. ACM, 2012.

[LAC12] Neal Lathia, Saniul Ahmed, and Licia Capra. Measuring the impact of opening the london shared bicycle scheme to casual users. *Transportation Research Part C: Emerging Technologies*, 22:88–102, 2012.

[LKAC12] Matthew Lave, Jan Kleissl, and Ery Arias-Castro. High-frequency irradiance fluctuations and geographic smoothing. *Solar Energy*, 86(8):2190 – 2199, 2012. ¡ce:title¿Progress in Solar Energy 3¡/ce:title¿.

[MBBE13] Sean Meyn, Prabir Barooah, Ana Busic, and Jordan Ehren. Ancillary service to the grid from deferrable loads: the case for intelligent pool pumps in florida. *ACE (GW)*, 100(50):50, 2013.

[PGTLB13] Miroslav Popovic, Peng Gao, Dan-Cristian Tomozei, and Jean-Yves Le Boudec. On the necessity of traffic shaping for pmu measurement data streams. In *Proc of Power and Energy Automation Conference*, 2013.

[PMN+09] Pierre Pinson, Henrik Madsen, Henrik Aa Nielsen, George Papaefthymiou, and Bernd Klöckl. From probabilistic forecasts to statistical scenarios of short-term wind power production. *Wind energy*, 12(1):51–62, 2009.

[SF04] Amer Shalaby and Ali Farhan. Prediction model of bus arrival and departure times using AVL and APC data. *Journal of Public Transportation*, 7(1):41–62, 2004.

[SHvH13] Jasper Schuijbroek, Robert Hampshire, and Willem-Jan van Hoeve. Inventory rebalancing and vehicle routing in bike sharing systems. Technical report, CMU, February 2013.

[Tri14] Mirco Tribastone. Efficient optimization of software performance models via parameter-space pruning. In *5th ACM/SPEC International Conference on Performance Engineering (ICPE'14)*, Dublin, March 2014. To appear.

[TT13] Max Tschaikowski and Mirco Tribastone. Insensitivity to service-time distributions for fluid queueing models. In *7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, 2013.

[VM11] P. Vogel and D.C. Mattfeld. Strategic and operational planning of bike-sharing systems by data mining: a case study. In *Proceedings of the Second international conference on Computational logistics*, ICCL'11, pages 127–141, Berlin, Heidelberg, 2011. Springer-Verlag.

[VS13] Andrea Vandin and Stefano Sebastio. Multivesta: Statistical model checking for discrete event simulators. In *7th International Conference on Performance Evaluation Methodologies and Tools (VALUETOOLS)*, 2013.

[WAW11] Wei Wang, John P. Attanucci, and Nigel H. M. Wilson. Bus passenger origin-destination estimation and related analyses using automated data collection systems. *Journal of Public Transportation*, 14(4), 2011.

[WNPK+11] G. Wang, M. Negrete-Pincetic, A. Kowli, E. Shafieepoorfard, S. Meyn, and U. Shanbhag. Dynamic competitive equilibria in electricity markets. In A. Chakrabortty and M. Illic, editors, *Control and Optimization Theory for Electric Smart Grids*. Springer, 2011.

[YGW+13] Yingxiang Yang, David Gerstle, Peter Widhalm, Dietmar Bauer, and Marta Gonzalez. The potential of low-frequency avl data for the monitoring and control of bus performance. In *Transportation Research Board 92nd Annual Meeting Compendium of Papers*, 2013.

[ZRW07] Jinhua Zhao, Adam Rahbee, and Nigel HM Wilson. Estimating a rail passenger trip origin-destination matrix using automatic data collection systems. *Computer-Aided Civil and Infrastructure Engineering*, 22(5):376–387, 2007.