MSc Dissertation Report

Probability Distribution of Wind Power during Peak Demand

Pablo Esteban Olmos Aguirre
0671810
August 2008

MSc Project Mission Statement

Effect of wind generation on network security standards

Student: Pablo E. Olmos Aguirre
Supervisor: Prof. Janusz Bialek & Dr. Chris Dent

Background

The current UK generation security standards were designed around thermal and hydro generation, where if a unit is working it can if required run at maximum output unless it is down for reparation or scheduled maintenance. For this scenario, it was established that around 20% reserve capacity was necessary. With the increasing penetration of wind generation, whose available output is determined by the weather conditions, the analysis underlying the security standard must change.

Objectives

This project will construct a probability distribution for the available wind generation in the UK based on wind data at sites around the country, and investigate what the appropriate capacity credit for wind generation should be in the generation security standard (i.e. what proportion of rated capacity should be assumed available when performing a security analysis.) The analysis will include the projected growth in penetration of wind production for the next 5 years.

The supervisor and student are satisfied that this project is suitable for performance and assessment in accordance with the guidelines of the course documentation.

Preliminary objectives

- Context and background reading of wind energy resource modelling
- Familiarisation with statistical tools and methods
- Familiarisation with the GB Security and Quality of Supply Standard

Signed

Pablo E. Olmos Aguirre

Prof. Janusz Bialek

Dr. Chris Dent

Date

Abstract

What proportion of the rated capacity of wind power should be assumed available when performing a security analysis? This report answers the question by providing the regional and GB-wide probability distributions of wind power over three time periods: longterm, winter and during peak demand. To obtain these distributions, a methodology is provided. The methodology takes a dataset of hourly wind speed observations from the Met Office, cleans the dataset and models the wind speed as power output. The GB distribution is calculated from a weighted average of the regional distributions. The weights realistically reflect where current and future wind generation is and will be. It was found that the peak demand capacity factor is better than the longterm, but worse than the winter capacity factors.

Declaration of Originality

I declare that this thesis is my original work except where stated.

Pablo Esteban Olmos Aguirre

Nomenclature

α Power law exponent

v Wind speed

z Height

z_r Reference height

10% peak demand The 90% percentile of peak demand

15% peak demand The 85% percentile of peak demand

20% peak demand The 80% percentile of peak demand

5% peak demand The 95% percentile of peak demand

ACS Average cold spell

BADC The British Atmospheric Data Centre

BERR Department for Business, Enterprise & Regulatory Reform

BWEA British Wind Energy Association

ea East Anglia

ee East England

em East Midlands

GB Great Britain

GB SQSS Great Britain Security and Quality of Supply Standard

ldn London

Met Office UK Meteorological Office

MIDAS Met Office Integrated Data Archive System

nee	North East England
nes	North East Scotland
ni	Northern Isles (i.e. Shetland)
nw	North Wales
nwe	North West England
nws	North West Scotland
se	South England
see	South East England
ss	South Scotland
sw	South Wales
swe	South West England
wm	West Midlands

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The GB Security and Quality of Supply Standard (GB SQSS) defines a coordinated set of criteria and methodologies for the planning and operation of the GB transmission system [13].

> "[The Standard] was established for a power system predominately [sic] supplied by conventional generation and has provided the basis for the development of an economic and efficient transmission system over the years.
>
> The amount of renewable generation (particularly wind generation) connected to the grid is increasing as a consequence of the government's aspirations to reduce greenhouse gas emissions from electricity generation. Many renewable generation sources are intermittent with characteristics significantly different from those of conventional generation. Due to the variability of these sources, their ability to contribute during times of peak demand is lower compared with their conventional counterparts, however, when it is windy, the wind powered generation places a higher value on access to transmission capacity." [15]

How does one determine the ability of wind farms to contribute during times of peak demand? Naturally, by the generating capacity they have. However, they cannot produce at full output all the time as they are bound to the winds variability. Their contribution is therefore expressed as a proportion of their generation capacity – their capacity factor.

1.1 Objective

The objective of the dissertation is to answer the question:

> What proportion of rated capacity should be assumed available when performing a security of electricity supply analysis?

The answer lies in the probability distribution for the power output from wind.

A methodology for modelling wind speed measurements as wind power capacity factors will be presented in this dissertation report. The outcomes of the model are probability distributions of wind power in three different time frames: longterm, winter and peak demand.

1.2 Scope

This project has developed a methodology to manipulate historic onshore wind speed data from the Met Office. The outcomes of the methodology are region- and GB-wide probability distributions of the capacity factor of wind power in the longterm, winter and peak demand time frames. The capacity factor is modelled as the hourly output of a simulated wind turbine.

Offshore wind speed data were not considered for this work. The temporal range of the data used is from the begining of 2001 until the end of 2007.

1.3 Structure

The next chapter makes a detailed description of the methodology and further justifies the importance of the chosen time frames. Chapter 3 presents the results and discusses them. The conclusions of this project and an outline of further work are reported in chapter 4. Finally, a technical appendix is included. The appendix briefly shows the actual implementation of the methodology. It is accompanied by a DVD that contains all the work of the project.

Chapter 2

Methodology

2.1 Introduction

This research was done using the wind speed data from the land surface observations of the Met Office Integrated Data Archive System (MIDAS) [22]. The dataset consists of hourly observations made by automated or manned weather stations across the country. The wind speed observation is the mean wind speed over 10 minutes measured from minute 40 to minute 50 of the hour [21].

The stations are organised into networks that observe different meteorological phenomena such as climate, wind, rainfall, and solar radiation [21]. Any given weather station can be part of one or many different observation networks. Furthermore, the data collected by the networks overlap, i.e. several networks may record wind speed. This feature results in multiple observations from the same weather station at the same date and time, but which correspond to different networks or domains.

The core purpose of the project is to obtain a probability distribution for the capacity factor of wind farms in Great Britain (GB). In what follows, the series of steps necessary to reach this goal are described.

2.2 Strategy

Given its size and location, it is reasonable to think that the wind resource varies widely in GB. Because of this, it is proposed to divide it in regions and then calculate individual probability distributions along with a *weighted*

Code	Region name
ea	East Anglia
ee	East England
em	East Midlands
ldn	London
nee	North East England
nes	North East Scotland
ni	Northern Isles (i.e. Shetland)
nw	North Wales
nwe	North West England
nws	North West Scotland
se	South England
see	South East England
ss	South Scotland
sw	South Wales
swe	South West England
wm	West Midlands

Table 2.1: Regions

average depending on the capacity in MW of existing and future wind farms in each region.

2.2.1 Regionalisation

The geographical division was based on the existing tariff zones defined by National Grid [12]. Because Scotland has one of the best wind resources in Europe [19], it was further divided to have better detail.

Sixteen regions were defined. They are listed in table 2.1 and their location on the map is shown in figure 2.1. The code is a simple abbreviation and will be used to refer to a region throughout the document.

Once the regions have been defined, the first step is to select the wind speed data by each region in order to extract meaningful information.

Data grouping – first attempt

The stations that reported observations of interest are those that have been operating from 2001, or before, up to now. Therefore, a selection of stations was made where the start date of operation was earlier than or equal to 2001-01-01 and the scheduled end date of operation was later than or equal

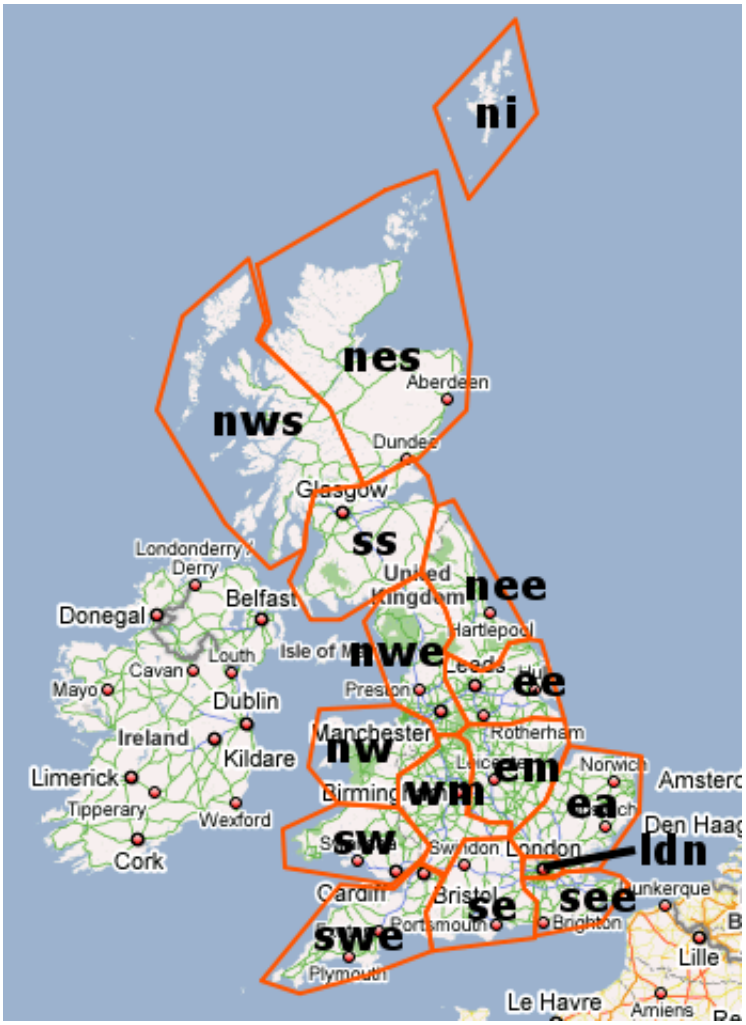

Figure 2.1: Defined regions for the study

to the present day (see sec. A.1.2). This yielded a list of stations which were then classified by regions according to their location (see sec. A.1.3). The result was sixteen sets of stations that were active from the beginning of 2001 to the end of 2007. From each list, 10 stations were randomly selected. The wind speed observations that were made by these stations can now be gathered.

Yet, it was found that this strategy would not work as the stations that *actually* recorded the observations in the dataset are an extremely small subset of the ones that are supposed to be active. A different approach is necessary.

Data grouping – second attempt

With the previous experience in place, a list of the stations that actually had recorded the observations in the wind speed dataset was compiled. This list was then sorted according to the regions (see sec. A.2.2). The stations grouped by regions can be found in table A.3. The number of stations for each region range from 1 to 14; because of this variability it was decided to keep all the stations in a region for the calculations rather than selecting a fixed number for each.

The framework to start working with subsets of observations from which regional information can be extracted is now in place.

2.2.2 Information extraction

The smallest bit of information that is sought is the capacity factor of hypothetical wind farms for any given region for *each hour* from the period 2001-01-01 00:00 hrs to 2007-12-31 23:00 hrs.

The capacity factor is the actual output of a generation unit over a period of time (e.g.500 MW h) divided by the rated capacity times the period of time [7]. The time period for an hourly capacity factor is 1 hour. This yields the following equation:

$$\frac{\text{output}}{\text{rated capacity}} = \text{capacity factor} \tag{2.1}$$

The power output of a wind turbine is determined by the wind speed at any given moment and every turbine has a characteristic curve that describes

this relationship. The wind turbine used for this study is the Bonus 2 MW[1]. The turbine's hub stands at 60 m from the ground. Its power curve is shown in figure 2.2. The cut-in and cut-out wind speeds mark the lower and higher bounds at which any turbine operates [18], for the Bonus it is 4 and 25 m/s respectively.

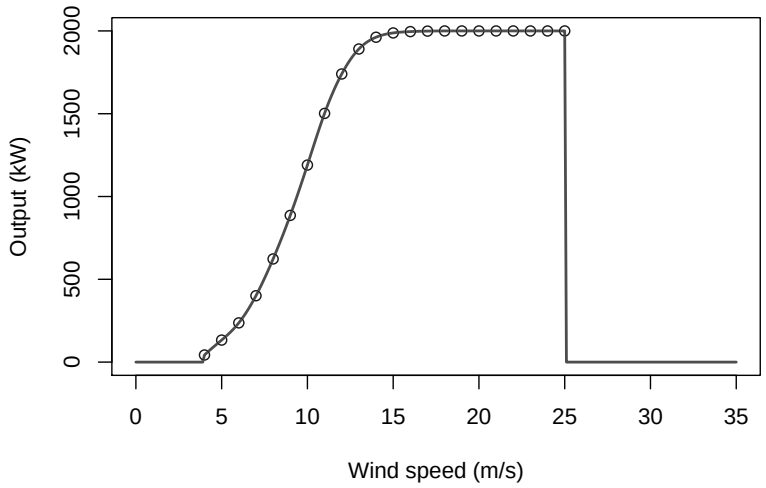

Figure 2.2: Bonus 2 MW power curve

The MIDAS wind speed data are in knots. A knot is equal to one nautical mile per hour and in turn is equal to $1852\,\mathrm{m}/3600\,\mathrm{s} = 0.51444444\,\mathrm{m/s}$ [9].

Furthermore, it is measured at 10 m, so a height extrapolation of the wind speed is necessary. This is achieved by applying the power law equation:

$$\frac{v_z}{v_{z_r}} = \left(\frac{z}{z_r}\right)^{\alpha} \tag{2.2}$$

where v_z is the wind speed at height z, v_{z_r} is the reference wind speed at height z_r and α is the power law exponent; it is common practice to use $\alpha = 1/7$ [18].

The hourly capacity factor for a region is the average of capacity factors at the same hour of the same day of the same month of the same year over

[1] This turbine was chosen because of the availability of the actual points defining the power curve. See [26]

the N weather stations in that region.

To summarise, these are the steps required to obtain the information we need:

1. Start with the wind speed in knots measured at 10 m
2. Convert to $^{m}/_{s}$ and apply power law to extrapolate to 60 m
3. Transform to capacity factors
4. Obtain hourly average over all the capacity factors in a region

Before these steps are taken, the datasets must become available for manipulation.

2.3 Data manipulation

The wind speed dataset for the period of 2001 to 2007 contains over 12 million records. The station dataset file contains about 5 thousand entries. Common spreadsheets are not fit to handle such a deluge of data.

2.3.1 Tools for data manipulation

The most practical way of filtering, manipulating and polling vast datasets is by loading them onto a database system. And yet, a database is not designed to be a serious tool for statistics, so a statistics package is needed to process and plot the data and render them into information.

For this purpose, MySQL* and R** were selected as database system and statistical tool, respectively. Both are opensource software and are freely available online for a myriad of different operating systems.

2.3.2 Integrity

The MIDAS Land Surface Observation records date from 1853 [21]. Any effort to collect and store such large amounts of data spanning such a long period is bound to meet data integrity problems. The MIDAS data is not an exception.

*http://mysql.com/
**http://www.r-project.org/

Station dataset

The chief problems found with the station list were:

- Different date formats. e.g. 2001-01-01, 1/1/2001, etc
- Latitude and longitude values are swapped or misplaced.
- The description of the data is missing and varies from version to version of the dataset

Some effort was put in cleaning up this data.

Observations dataset

Data integrity problems can arise because either the measuring instrument recorded a bad value, a value in non-standard units or none at all. Also, it could have been incorrectly typed-in. The MIDAS data are subject to certain quality control procedures to mitigate these problems. The quality control process tags and modifies (without losing the original data) the records. Individual records are tagged and can be duplicated according to their status in the process and the changes made.

Because this is the core source of information for this project, much effort was put in *filtering out* the corrupt data to use only the reliable values.

2.3.3 Loading

The MIDAS Land Surface Observation Stations Data can be used and downloaded for academic purposes from the British Atmospheric Data Centre (BADC) at http://badc.nerc.ac.uk. The dataset is split in yearly text files. The structure of each file is that of a table where the columns are separated by the comma character ",".

Within the file, the columns of interest are the observation date and time, the station identifier, the status of the record, the wind speed and the quality control flags on the wind speed value.

MySQL can interpret and load different data types (i.e. integer, decimal, date, etc.) from text. After close inspection of the types of data present in the dataset, a database structure or 'schema' was designed to contain the data.

The dataset was loaded but many value interpretation warnings and truncated values occurred during the process. It is time consuming to attempt to find and correct each error from the text files themselves. Instead, a simplified (where the type of all the columns are text) schema was used (see sec. A.2.1). Filtering out the values that caused the parsing warnings will be more flexible once the data are loaded in the database.

2.3.4 Clean up

The first step in cleaning up the data is to use the quality control (QC) information for each record in the dataset. There are two bits of information, the record status indicator and the q value.

Record status indicator & Version number

The record status indicator "is used to describe the current stage in the life of a particular record, from creation to deletion" [24]. A document provided to BADC by the Met Office describes the values and meanings of the status indicator. Table A.1 shows the different record states and their occurrence count (see sec. A.2.4).

According to the description of the status, it is best to avoid records which are in process of quality control or have been superceded by a new revised version. The majority (89%) of records are those with who have ended the process, therefore, these will be considered reliable.

q value

The next step is to look at the q value. Each record has one associated to it to "show the progress of the data through the quality control" [23]. "This [q] attribute is a five digit number [...], where each digit describes one aspect of the quality of a meteorological element" [23]. Table A.1 presents the occurring values and count (see sec. A.2.3).

The records that are flagged as "Observed and suspect" should obviously be avoided. Whereas there are others where it is not immediately clear from the description if they are reliable or not. For the sake of simplicity, only those with flags indicating that quality control has been performed were considered as reliable; these constitute 85.5% of the dataset.

Value precision

In [21] it is noted that the precision of the wind speed observation is 1 knot. With this in mind, any decimal value should be considered an invalid observation and thus filtered out. 96.9% of the observations in the dataset have integer wind speeds (see sec. A.2.5); these will be considered reliable.

Observations domain

Theoretically, a single station should have $(6 \cdot 24 \cdot 365) + (24 \cdot 366) = 61,344$[†] observations in the period from 2001 to 2007. It was found, however, that there were stations that had over 84,000 observations. This happens because, as explained above, some stations belong to several domains (observation networks) and thus reported a wind speed for each domain. Table A.2 shows the different domains and their occurrence count.

The majority (73%) reports the Hourly Climate Message (HCM), so it is preferable to use this subset of observations.

To summarise, three steps must be taken to filter out the corrupt or duplicated data. Select records that:

- Belong to the HCM domain
- Have ended the quality control process (indicated by the record status and q value)
- Have an integer value as wind speed

This filtering process yields 9,137,903 reliable records (see sec. A.2.7). They represent 72.5% of the original dataset which has 12,595,404 records.

2.3.5 Conversion to m/s and extrapolation to 60 m

Now that a subset of clean data has been created, the transformation from knots to m/s and from 10 m to 60 m can be made (see sec. A.2.8). As noted above, the conversion factor from knots to m/s is 0.514444m/s:

[†]2004 is a leap year

$$v_{10}{}^{\mathrm{m/s}} = 0.514444\,{}^{\mathrm{m/s}}\, v_{10}\,{}^{\mathrm{nm/h}} \tag{2.3}$$

Once converted, the power law equation is used to extrapolate the wind speed to a height of 60 m:

$$\begin{aligned} \frac{v_z}{v_{z_r}} &= \left(\frac{z}{z_r}\right)^{\alpha} \\ v_z &= v_{z_r}\left(\frac{z}{z_r}\right)^{\alpha} \\ v_{60}{}^{\mathrm{m/s}} &= (0.514444)\; v_{10}\,\mathrm{kn}\left(\frac{60}{10}\right)^{1/7} \\ v_{60}{}^{\mathrm{m/s}} &= (0.664506)\, v_{10}\,\mathrm{kn} \end{aligned} \tag{2.4}$$

The hourly average for each region will be calculated after the transformation to capacity factor. The next section looks into using the hypothetical wind turbine to determine power output and capacity factor.

2.3.6 Transformation to capacity factor and aggregation

Let Ω be the set of all observations, σ the set of all stations that have recorded observations in Ω, and σ_r, $r \in \{ea, ee, ...\}$ a subset of σ defining a region (see table 2.1). An observation v_{is} is the wind speed indexed in terms of the date-time i and the station $s \in \sigma_r$ that recorded it.

Transformation to Capacity Factor

The points of the power curve (see figure 2.2) for the Bonus 2 MW were obtained from [26]. To make the calculations of the power output, a function that fits these points was approximated (see sec. A.3.1). Let $p(v)$ denote this function. The cut-in and cut-out speeds of the Bonus 2 MW are 4 and 25 $^{\mathrm{m/s}}$. This creates two discontinuities in the function[‡] p. So we define a new function $\hat{p}(v)$ as:

[‡]the original function is continuous and returns unsuitable values for windspeeds below 4 and above 25 $^{\mathrm{m/s}}$

$$\hat{p}(v) = \begin{cases} 0 & \text{if } v < 4 \\ 0 & \text{if } v > 25 \\ p(v) & \text{otherwise} \end{cases} \tag{2.5}$$

The wind speed is transformed to power using $\hat{p}(v)$ and then to a capacity factor κ by dividing it by 2000 (the rated output of the Bonus turbine in kW) and scaling it by a factor of 100 (equation 2.6). This function is applied to the whole set of observations obtaining hourly capacity factors κ_{is} (see sec. A.3.3).

$$\kappa_{is} = \frac{\hat{p}(\bar{v}_{is})\,\text{kW}}{2000\,\text{kW}} 100\% \tag{2.6}$$

Regional hourly average

The hourly average capacity factor for each region $\bar{\kappa}_{i\sigma_r}$ is then defined as:

$$\forall r,\ \forall i,\ \ \bar{\kappa}_{i\sigma_r} = \frac{1}{N} \sum_{s \in \sigma_r} \kappa_{is} \tag{2.7}$$

Where N is the number of stations in σ_r.

National weighted average

A straight or simple average across all the regions would hide any high or low capacity factor values. To avoid this, a weighted average is used. It is generically defined [6] as:

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \tag{2.8}$$

Where $[x_1, x_2, \ldots, x_n]$ is the dataset and $[w_1, w_2, \ldots, w_n]$ are the weights.

The weights in table 2.2 are based on the aggregated MW capacity of the operational, in construction, consented and in planning wind farms in each region (see sec. A.3.2). The data for the wind farms were obtained from the BWEA website [1, 3, 2, 4] and later classified by regions. For a visual representation of the weights see figure 2.3.

In the chapter 3 the results for the regional and GB capacity factors are provided. As a means of comparison, an arithmetic average is shown.

However, the results obtained from the above process are too broad and

Region	Weight
ea	331.025
ee	482.300
em	423.325
ldn	3.600
nee	675.160
nes	2319.300
ni	3.680
nw	1599.350
nwe	1389.700
nws	915.325
se	95.900
see	90.850
ss	4394.650
sw	517.830
swe	264.925
wm	30.400

Table 2.2: Regional weights

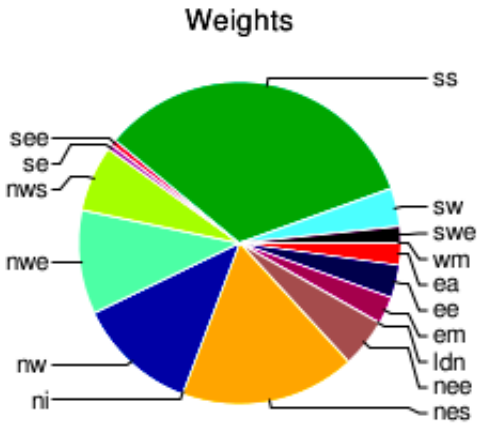

Figure 2.3: Regional weights

mean little in terms of the Security and Quality of Supply Standard. The standard refers to the performance of all installed capacity during *peak demand*. So a look at the wind power capacity factors during the set of hours where peak demand occurs is vital.

2.4 Capacity factors during peak demand

The *crux* of this study is the regional and GB capacity factors during hours when demand for electricity are highest – peak demand.

Peak demand is an estimation of demand during winter under the condition of an Average Cold Spell (ACS) [13]. The ACS is "a particular combination of weather elements which give rise to a level of peak demand [...] which has a 50% chance of being exceeded as a result of weather variation alone" [13].

National Grid has a plan to meet ACS Peak Demand when it occurs (refered to as a 'Planned transfer condition' [13]). It essentially consists of calling to operation or scaling the output of *preselected* generators whose aggregated capacity equals that of peak demand plus a margin [13]. The selection of generators for the planned transfer condition is a three step process:

- "Ranking the relevant generating units in order of their relative likelihood of operation at peak;
- Identifying which plant is most likely to be contributing towards meeting the peak demand; and finally
- Applying the straight scaling technique". [14]

The pertinent steps of the process for this study are those where the generators are selected on the basis on their likelihood to contribute to the goal. To determine the likelihood, the measure of the availability is used [14]. The availability factor is the ratio between the amount of time that a generator is *able to operate* to the time period of winter [14, 18].

To analyse the contribution of generators during winter peak, the winter capacity factor becomes relevant. For wind farms, 'to be able to operate' not only means to be in proper working order throughout the year but also that the wind speed is sufficient to generate power (reflected by the capacity

factor). Thus, to make it comparable, wind generation is converted "to an equivalent thermal unit that has the same availability at ACS peak" [15]. The availability of a conventional thermal unit is assumed to be be 90% [13]. So the contribution of a wind farm with, say, 36% capacity factor is 0.36/0.9 = 0.4; for example, a 100MW wind farm would be treated as a 40MW thermal unit [15].

Once ordered according to their availability, the lower ranking generators are progressively removed, until the aggregated capacity exceeds the amount of the ACS peak demand by 20% [14]. This aggregated capacity is formally referred to as 'Plant Margin' [13, 14].

This margin is necessary for security of electricity supply [14]. Because the capacity factor of wind generation during winter peak is used in the calculations of the plant margin, it is essential to extend the methodology to determine the hours at which winter peak demand occurs and look at the capacity factor of wind in that time frame. In what follows, a description of how these hours were determined is made.

2.4.1 Demand Data

Historic demand data is available from the National Grid website [11]. The data used for comparison is the Total Gross System Demand which is taken in 30 minute intervals and spans from April 2001 until the current date. The range considered for this study is from April 2001 to the end of December 2007.

2.4.2 Data manipulation

The demand data show an increase from year to year. Naturally, peak demand during 2007 will be greater than during 2001. To work around this, the data are split in years taken from July 1 00:00 hrs of a year to June 30 23:00 hrs of the next year. The maximum value in that time range is taken as 100% and then rest are normalised to it.

The information of interest is the actual date and time of the data, not the normalized demand value. So the hours that fall within the 80%, 85%, 90% and 95% percentiles are selected (or conversely the top 20%, 15%, 10% and 5%). The range from 20 to 5% is considered for comparison purposes.

All the peak demand hours from each 'demand year' are gathered in a

Range	Hour count
2001-2002	46
2002-2003	42
2003-2004	82
2004-2005	116
2005-2006	120
2006-2007	97
Total	503

Table 2.3: Number of demand hours in the top 5% percentile of the peak

list. Table 2.3 shows a summary of how many hours were selected in each year.

There are 118,358 demand records in the period from April 2001 to the end of 2007, two for each hour. Half of them are left after taking the maximum value of each hour, that is, 59,179. The 503 peak hours represent 0.84% of the latter total.

To obtain the regional and GB capacity factors during peak demand, one must select the capacity factors whose date-time index *match* the peak hours. In other terms, let Π be the set of peak hours j. Where j is a date-time. If K is the set of all capacity factors $\kappa_{i\sigma_r}$ then we seek a subset K' where $\kappa_{j\sigma_r} \in K | j \in \Pi$.

It was found that not all the peak hours could be matched to the capacity factor index. It is assumed that this is because of observation gaps in the MIDAS dataset. So, a slightly smaller (497 hours out of 503) list is obtained. This is the dataset of interest.

The above selection criteria were chosen because, at the time of defining the methodology, historic ACS peak demand forecasts were not obtained. At the time of writing, however, the forecasted ACS peak demand was found to be $61.4GW$ [14]. The criteria yield hours where demand is around 98% of $61.4GW$; therefore, it can be considered appropriate.

The next chapter shows the results of the aforementioned methodology and calculations.

Chapter 3

Results

3.1 Overview

Probability distributions of wind power for each region and for GB (average over all regions) are presented. The distributions were calculated with the hourly capacity factors from three time ranges: long term, winter, and during peak demand.

The distributions are represented as histograms with a bin width of 4%. Each figure has the mean capacity factor marked with a dashed vertical line. Features of the distributions are described by making reference to the range (e.g. 4–8%) of capacity factors where they occur.

In general, the ranking of regional capacity factors was fairly similar across all three time ranges. Shetland is always on the top and either London or the West Midlands at the bottom. To be fair, because Shetland is not connected to the grid [17], it should not be part of the ranking. This leaves the north west of Scotland as the top region for wind power.

London was included in all the calculations because it is defined as a tariff zone by the National Grid. However, the value of the information extracted for it is questionable. This is because in that region only one station had recorded observations (an effect of its small area). Furthermore, the bin width of 4% for the histogram seems too small as the resulting graph is considerably jagged making conclusions somewhat uncertain.

London and Shetland are the two lowest contributors to the weighted average. This is an appropriate effect of the process to obtain the weights. Although appropriate, it may not be satisfactory, for the reasons explained

above.

Most regional probability distributions have one maximum at 0–4% capacity factor. The exceptions are Shetland and North West Scotland. Shetland has two maxima in each extreme, one at 0–4% and the other at 96–100% capacity factor. North West Scotland has a very low maximum at around 4–8%, and has the additional characteristic of being fairly flat or uniform across the horizontal axis. The other regions behave somewhat like an exponential. As the distributions descend the ranks, the percentage of hour for 0–4% capacity factor increases. This is a result of worse wind resource.

All GB distributions have a maximum at around 4–16% capacity factor except "5% peak" which is at 0–4%. After the maxima, the distributions decrease reasonably fast, yet, the tails are "fat". All weighted averages are slightly better than the unweighted ones. Also, the percentage of hours for maximum output (96–100% capacity factor) is zero for all of the GB distributions.

3.2 Long term average

The long term average is calculated over the full range of the dataset, that is, from the beginning of 2001 to the end of 2007. In this date range there are 61,344 hours. The resulting dataset of the methodology has 60,167 hourly observations (98.08% of the total).

3.2.1 Regional averages

The regional average capacity factors range from a maximum of 36.21% for North West Scotland to a minimum of 11.03% for the West Midlands. Figure 3.1 shows the regional longterm distributions.

3.2.2 GB average

The longterm average (weighted) capacity factor of 24.26% found in this work for GB is slightly lower than the averages published by BERR [8] and Sinden [25] but higher than the one from the SQSS Review Group [15]. The BERR average for the years 2002–2007 is 27.65%[1], the one for Sinden is 30%, the one for the SQSS Review Group is 22%. The unweighted average is marginally

[1] The longterm average is calculated from published annual averages

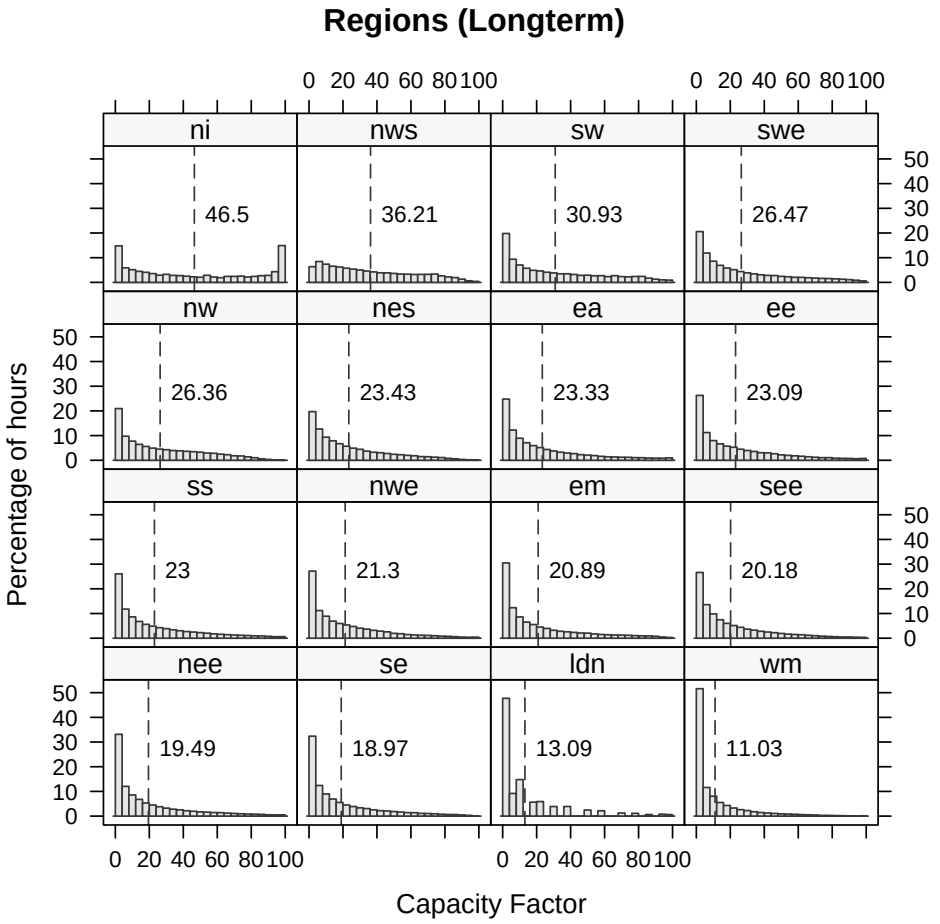

Figure 3.1: Long term regional histograms, ordered by means

worse at 24.02%. As mentioned earlier, the unweighted distribution peaks at 8–12% capacity factor whereas the weighted peaks at 4–8% (figure 3.2).

3.3 Winter average

The winter average is calculated over the six full natural winters of the dataset. The winters are taken from the day of the winter solstice to the day of the vernal (spring) equinox (for the northern hemisphere) [20], that is, from December 21 00:00 hrs to March 20 23:00 of every year. In this date range there are 12,840 hours. The resulting dataset from the methodology has 12,804 hourly observations (99.71% of the total).

The SQSS document, Review for Onshore Intermittent Generation, calculates winter peak capacity factors from the beginning of December to the end of February [15], whereas the ACS peak demand is calculated from late-October to late-March[2] [14]. For comparison purposes, all three winters, unweighted and weighted, are shown in figure 3.4.

A particular characteristic of the regional and GB averages is that the capacity factors are sustained for a larger percentage of hours. This shows that the wind resource is more plentiful during winter.

3.3.1 Regional averages

The regional average capacity factors for winter range from a maximum of 45.62 for North West Scotland to a minimum of 17.26 for the West Midlands. It is worth noting that all winter averages are well above the longterm averages. Moreover, South Wales, South West England and North Wales also show a roughly uniform distribution in the centre. Figure 3.3 shows the regional distributions.

3.3.2 GB average

The weighted winter average is 32.62% capacity factor (top right of figure 3.4) and peaks at 4–8%. The unweighted average is lower at 32.43% (top left of figure 3.4) peaks at 12–16%.

The graphs in the middle left and right show the effect of taking different time ranges for the definition of winter. In general, the shape of the distribu-

[2] This was interpreted as a range from the last day of October to the last day of March

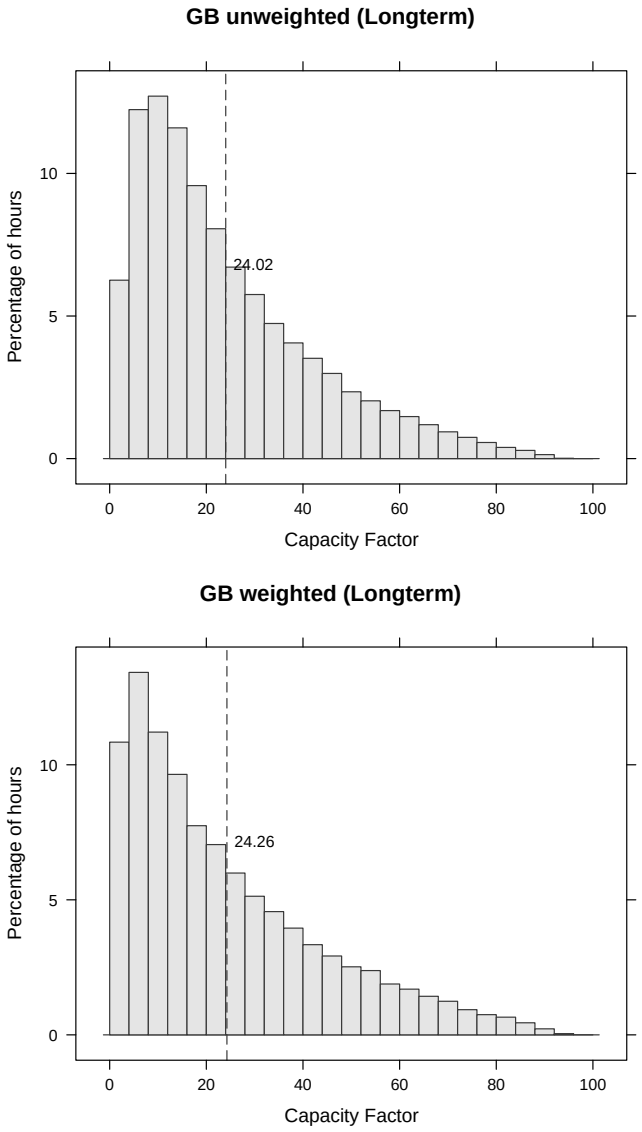

Figure 3.2: Long term arithmetic and weighted average

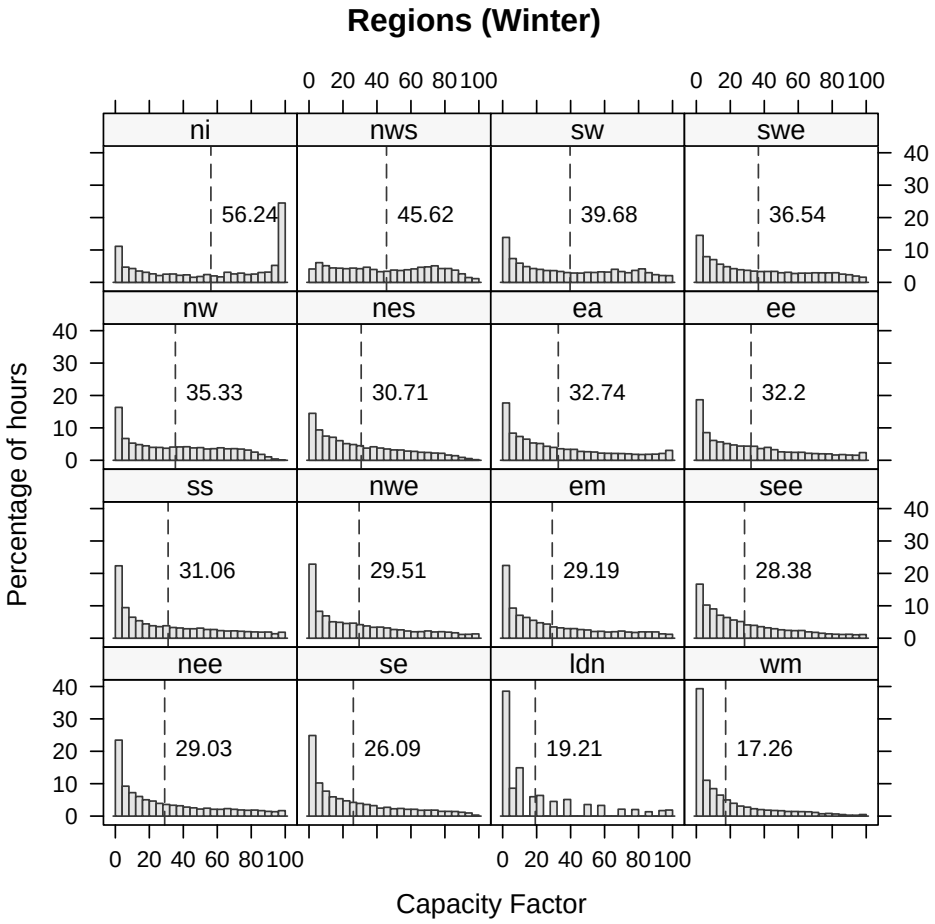

Figure 3.3: Natural winter regional histograms, ordered by means

Percentile	Hour count
20%	9,175
15%	5,051
10%	1,844
5%	503

Table 3.1: Number of demand hours in each percentile

tions does not vary, whereas the average capacity factor does. The natural winter average is the best, and the late October to late March is the worst.

3.4 Peak demand average

As described in section 2.4, the peak demand hours are determined by normalising the demand to its maximum value from intervals of July–June across the seven years of the demand dataset and then collecting the hours at the top 20, 15, 10 and 5% percentiles. Table 3.1 shows how many hours were considered for each one.

The main focus is on the 5% percentile, but a comparison with the others is also shown.

3.4.1 Regional averages

The regional average capacity factors for 5% peak demand are almost as high as those of winter. They range from a maximum of 41.29% for North West Scotland to a minimum of 11.3% for the West Midlands. Figure 3.5 shows the regional distributions.

3.4.2 GB average

The weighted 5% peak demand average is 25.99% and has its maximum at 0–4% (bottom left of figure 3.7), whereas the unweighted average is 25.48% with maximum at 12–16% capacity factor (bottom left of figure 3.6). It is worth noting that at peak demand, the weighted capacity factor is better than the longterm but considerably worse than winter.

As a comparison on how important the definition of 'peak demand' is, figure 3.7 shows the weighted distributions of 20, 15, 10 and 5% peak demand. The average capacity factor starts at 31.08% for 20% peak to 25.99% for 5%

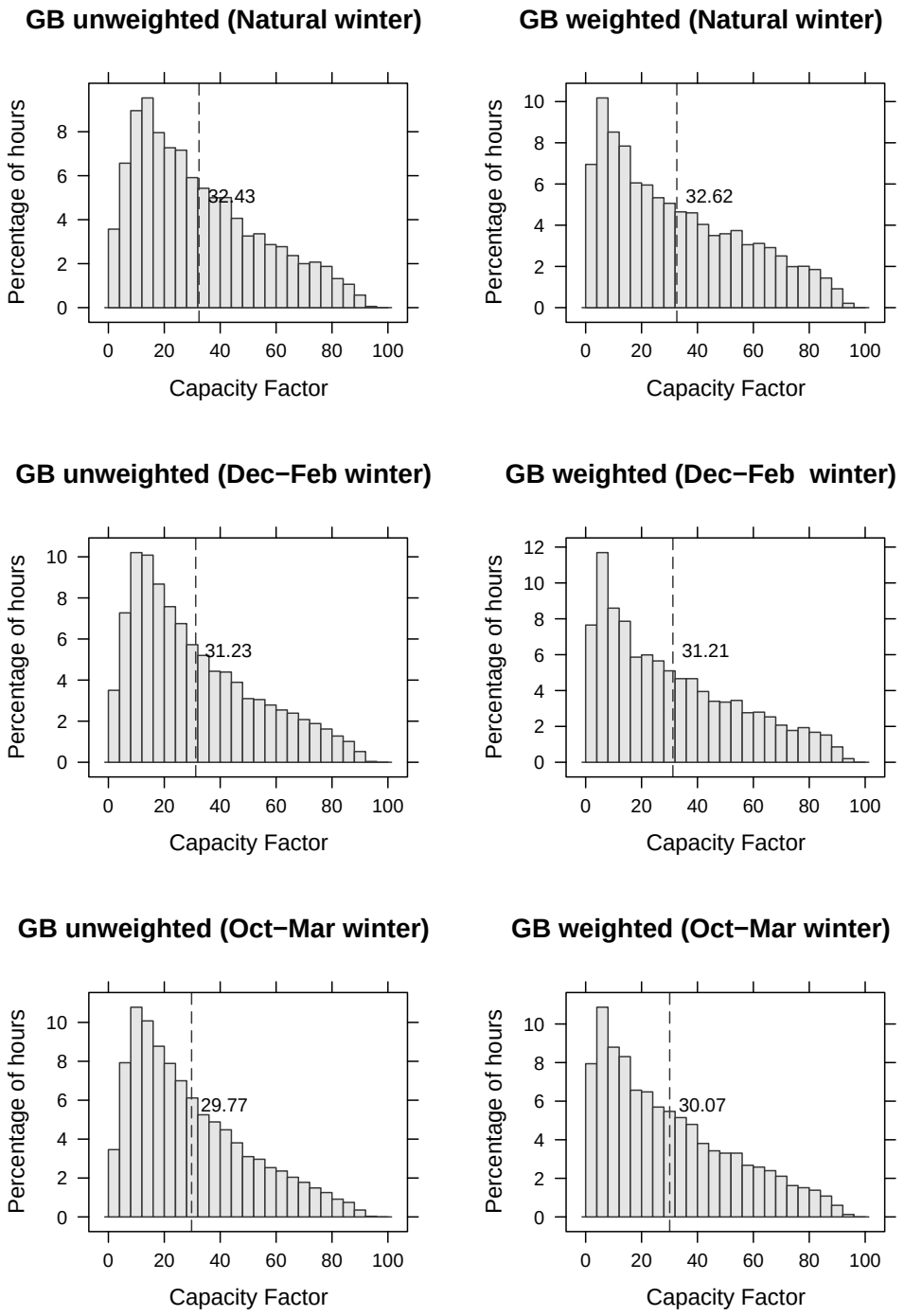

Figure 3.4: Comparison of the effect of the definition of winter to the unweighted and weighted distributions

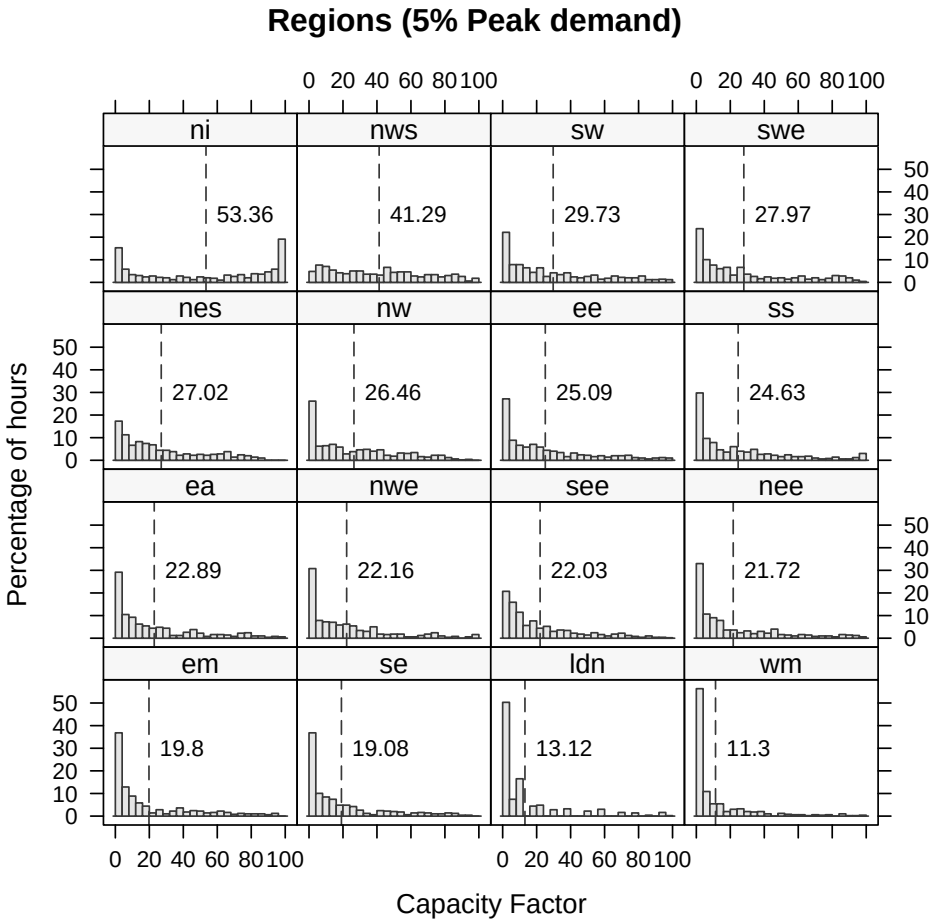

Figure 3.5: Peak demand regional distributions, ordered by means

peak. There is a 5.9 difference when you vary what peak demand is from 20 to 5%.

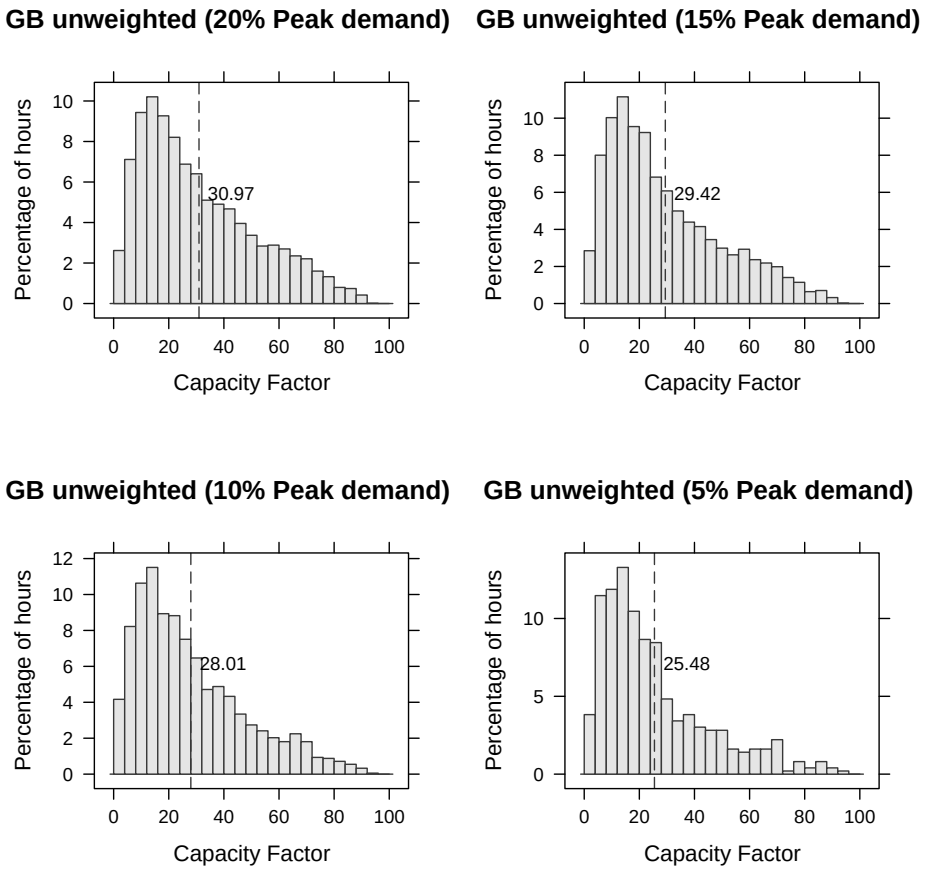

Figure 3.6: Comparison of distributions at different peak demand percentiles (Arithmetic average)

3.5 Summary

A summary of the GB distributions is presented in table 3.2. The Most Likely (maximum of the distribution) and the Mean capacity factors are shown. The mean is annotated by its estimated probability.

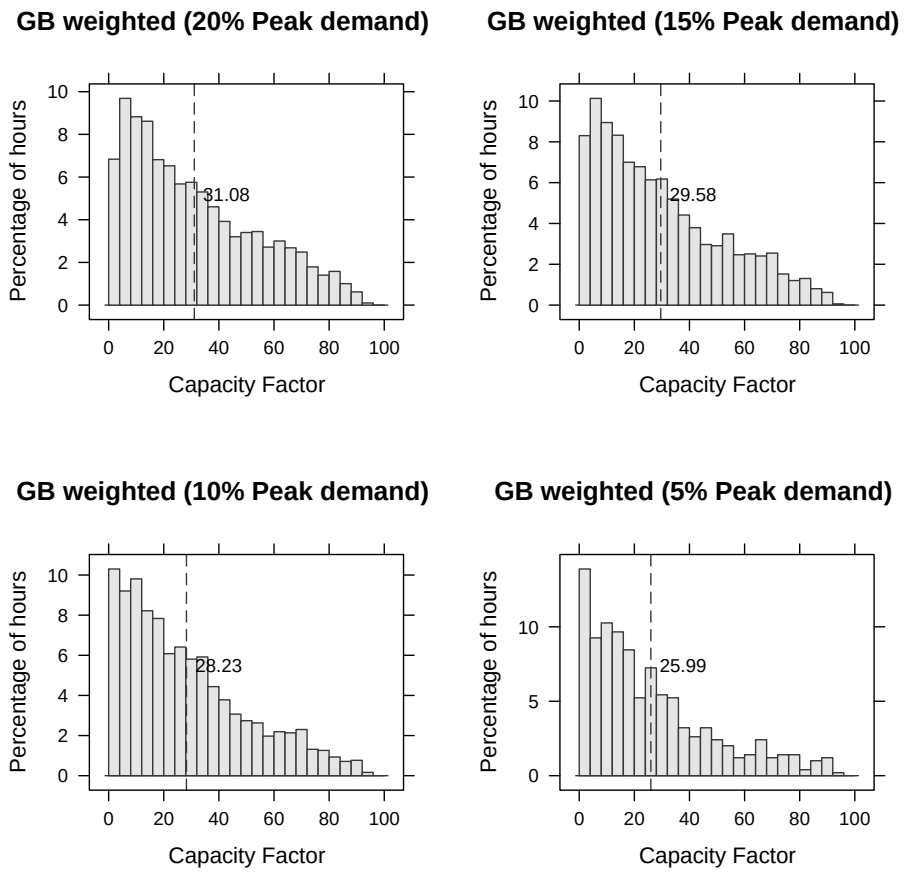

Figure 3.7: Comparison of distributions at different peak demand percentiles (Weighted average)

		Most likely	Mean
Longterm	u	8–12%	$24.0\% \approx 7\%$ of hours
	w	4–8%	$24.3\% \approx 6\%$ of hours
Natural Winter	u	12–16%	$32.4\% \approx 5.5\%$ of hours
	w	4–8%	$32.6\% \approx 4.5\%$ of hours
Dec–Feb Winter	u	8–16%	$31.2\% \approx 5.75\%$ of hours
	w	4–8%	$31.2\% \approx 5\%$ of hours
Oct–Mar Winter	u	8–12%	$29.8\% \approx 6\%$ of hours
	w	4–8%	$29.8\% \approx 5.75\%$ of hours
20% Peak demand	u	4–8%	$30.1\% \approx 6.25\%$ of hours
	w	4–8%	$31.1\% \approx 5.75\%$ of hours
15% Peak demand	u	0–4%	$29.4\% \approx 6\%$ of hours
	w	4–8%	$29.6\% \approx 6\%$ of hours
10% Peak demand	u	4–8%	$28.0\% \approx 6.25\%$ of hours
	w	0–4%	$28.2\% \approx 5.75\%$ of hours
5% Peak demand	u	0–4%	$25.5\% \approx 7.75\%$ of hours
	w	0–4%	$26.0\% \approx 6\%$ of hours

Table 3.2: Summary of results. ('most likely' and 'mean' are capacity factors; u and w refer to *unweighted* and *weighted*)

Chapter 4

Conclusions

In this dissertation a methodology for modelling wind speed measurements as wind power capacity factors was presented. The outcomes of the model are probability distributions of wind power in different time frames: longterm, winter and peak demand. These distributions provide the answer to the question posed at beginning of the dissertation.

The conclusions reached in this dissertation are detailed below.

- Using a weighted average to obtain a GB capacity factor improves its value. This improvement is not "staged" as the weights reflect where current and future wind generation is and will be.
- Considering different time ranges for winter notably affects the average capacity factor in that period. Consequently, this will affect the contributions of wind farms towards the Planned Transfer Condition and the amount of Plant Margin.
- The average capacity factors along the top percentiles of peak demand also show a considerable change (for worse) in values. The drop of the capacity factor from the top 20th to the top 5th percentile is of $31.1 - 26 = 5.1$, which is a significant amount.
- The 5% peak demand capacity factor is better than the longterm, but worse than the winter capacity factors. Nonetheless, the average capacity factor at 5% peak demand shows that wind power can still have a valuable contribution in times when demand is highest.

- During the development of the methodology, it was found that the order of the steps of transforming to hourly capacity factors and aggregating regional hourly averages is important (section 2.3.6). The consequence of inverting the steps was that the peaks of all distributions are shifted from around 4–16% to 0–4% capacity factor.

Limitations

An inherent limitation of any study similar to this one is the model itself. *Real* probability distributions can be obtained from actual wind farm output data, rather than wind speeds from weather stations. "The best way to understand the relationship between wind power availability and electricity demand would be to analyse production data or wind speeds from operational windfarms" [15]. Currently, these data are scarce and not easily available for all operating wind farms in the UK. Finally, the value used for the power law exponent α is a simplification as it corresponds to wind flow over flat planes [18].

The time frame for this study – 2001 to 2007 – is another limitation; longer periods are preferred to observe trends. The main reason for this time frame is the availability of historic demand data from National Grid (from April 2001 to the current date, see ref. [11]).

The resulting probability distributions for wind power in this study do not account for offshore wind generation. This is because "wind data at offshore locations are spatially and temporally sparse and of variable quality" [27].

Future work

The longterm average obtained in this study is considerably low compared to the longterm average published by BERR [8]. A scaling step is included in the methodology in the work published by Sinden [25]. This step scales the wind speed data to ensure a match of the modelled longterm average with the official average. A similar step for this methodology should be considered for future work.

Althought the regions used in this study are based on the tariff zones from National Grid, it was mentioned that a more appropriate division could be made. The new division would exclude Shetland and merge London to any

of its original surrounding regions.

Now that a probability distribution for wind has been determined, it should be fed into the calculations of the Planned Transfer Condition and the Plant Margin.

Accounting for offshore wind is a natural way of extending this work, perhaps by using data obtained from the POWER methodology described in ref. [27].

Acknowledgements

This dissertation would not have been possible without the kind help and support from:

my supervisors, Chris, Janusz and Colin
my girlfriend Rasa
my friends, Alejandra, Annie, Farlane, Jeff, Jo, Lauren, Niall, Owen, Peter, Ramón, Riccardo, Robin, Rodolfo and Siddhu
my parents, for all the love that God has given me through them
my brothers, Teresa, Luis and Gabriel
Deo gratias

I would like to thank them warmly for all they have done.

A special acknowledgement to the British Atmospheric Data Centre and the Met Office.

"The problems of the world cannot possibly be solved by sceptics or cynics whose horizons are limited by the obvious realities. We need men who can dream of things that never were."
John F. Kennedy

Bibliography

[1] British Wind Energy Association. Ukwed consented projects. http://www.bwea.com/ukwed/consented.asp. Accessed August 2008.

[2] British Wind Energy Association. Ukwed operational wind farms. http://www.bwea.com/ukwed/operational.asp. Accessed August 2008.

[3] British Wind Energy Association. UKWED projects in planning. http://www.bwea.com/ukwed/planning.asp. Accessed August 2008.

[4] British Wind Energy Association. Ukwed wind farms currently under construction. http://www.bwea.com/ukwed/construction.asp. Accessed August 2008.

[5] Paul Bourke. Determining if a point lies on the interior of a polygon. http://local.wasp.uwa.edu.au/ pbourke/geometry/insidepoly/. Accessed July 2008.

[6] G.M. Clarke and D. Cooke. *A basic course in statistics*. Hodder Arnold, 1998.

[7] U.S. Nuclear Regulatory Commission. Capacity factor (net). http://www.nrc.gov/reading-rm/basic-ref/glossary/capacity-factor-net.html. Accessed August 2008.

[8] Enterprise & Regulatory Reform Department for Business. Digest of united kingdom energy statistics 2007. http://www.berr.gov.uk/energy/statistics/publications/dukes/page39771.html. URN No: 07/87.

[9] Bureau International des Poids et Mesures. Table 8. other non-si units. http://www.bipm.org/en/si/si_brochure/chapter4/table8.html. Accessed August 2008.

[10] Google Code FAQ. What is kml? http://code.google.com/support/bin/answer.py?answer=56547&topic=10032. Accessed August 2008.

[11] National Grid. Demand data. http://www.nationalgrid.com/uk/Electricity/Data/Demand+Data/. Accessed July 2008.

[12] National Grid. GB generation use of system tariff zones 2007/08. www.nationalgrid.com/NR/rdonlyres/0D245D53-50A6-4F37-9EC9-94E76257C719/14719/GenerationZones2007_08.pdf. Accessed August 2008.

[13] National Grid. GB security and quality of supply standard. http://www.nationalgrid.com/uk/Electricity/Codes/gbsqsscode/DocLibrary/, September 2004. Version 1.0.

[14] National Grid. Gb seven year statement 2007. http://www.nationalgrid.com/uk/Electricity/SYS/archive/sys07, 2007.

[15] GB SQSS Review Group. Review for onshore intermittent generation. http://www.nationalgrid.com/uk/Electricity/Codes/gbsqsscode/DocLibrary/, January 2008. GSR001.

[16] Microsoft Help and Support. How to convert degrees/minutes/seconds angles to or from decimal angles in excel 2000. http://support.microsoft.com/kb/213449. Accessed August 2008.

[17] Highlands and Islands Enterprise. Assessment of national grid connection options for the scottish islands. http://www.hie.co.uk/HIE-economic-reports-2007/tnei-grid-study-june-07.pdf, March 2007. GSR001.

[18] J.G. McGowan J.F. Manwell and A.L. Rogers. *Wind Energy Explained.* John Wiley & Sons, 2002.

[19] University of Edinburgh. Matching renewable electricity generation with demand. Technical report, Scottish Executive, February 2006. http://www.scotland.gov.uk/Publications/2006/04/24110728/0.

[20] The American Heritage® Dictionary of the English Language. Winter. http://www.bartleby.com/61/6/W0180600.html. Fourth Edition, 2000. Accessed August 2008.

[21] UK Meteorological Office. Met office surface data users guide. http://badc.nerc.ac.uk/data/ukmo-midas/ukmo_guide.html. Accessed July 2008.

[22] UK Meteorological Office. Midas land surface stations data (1853-current). http://badc.nerc.ac.uk/data/ukmo-midas. British Atmospheric Data Centre, Accessed June 2008.

[23] UK Meteorological Office. Quality control information - met element name q (_q) and met element name j (_j). http://badc.nerc.ac.uk/browse/badc/ukmo-midas/metadata/doc/QC_J_flags.html. Accessed August 2008.

[24] UK Meteorological Office. State indicators. http://badc.nerc.ac.uk/browse/badc/ukmo-midas/metadata/doc/state_indicators.html. Accessed August 2008.

[25] Graham Sinden. Characteristics of the uk wind resource: Long-term patterns and relationship to electricity demand. *Energy Policy*, 35(1):112–127, January 2007. doi:10.1016/j.enpol.2005.10.003.

[26] WAsP – the Wind Atlas Analysis and Application Program. Power curves download page – sampletestfiles.zip. http://www.wasp.dk/Download/PowerCurves.html. Accessed August 2008.

[27] G. M. Watson; J. A. Halliday; J. P. Palutikof; T. Holt; R. J. Barthelmie; J. P. Coelingh; L. Folkerts; E. J. Van Zuylen. Power – a methodology for predicting offshore wind energy resources. www.eru.rl.ac.uk/POWER_project/bwea21_4.pdf.

Appendix A

Technical Appendix

The appendix provides all the technical steps executed to achieve the results. Ranging from the database schemas, queries, to the code to produce the plots. All code is annotated so that reproduction is possible.

The programming languages used for this project were SQL, R, BeanShell and Visual Basic.

As a guide to programming language lingo the following *names* are given. Any consecutive string of text characters will simply be refered to as 'string'.

This appendix tries to follow the same flow as chapter 2.

A.1 Regionalisation and data grouping

A.1.1 Station table definition

Before being able to select the stations, the data must be loaded onto the database. Because the metainformation for the table structure is missing, column names were guessed or left as unknown variables. Such names as *alpha, bravo, charlie*, etc. come from the NATO phonetic alphabet. "Whisky" is the name of the database and "stations" that of the table.

```
CREATE TABLE  `whisky`.`stations` (
  `src_id` int(11) NOT NULL,
  `src_name` varchar(64) default NULL,
  `lat` float default NULL,
  `lon` float default NULL,
  `loc_` varchar(16) default NULL,
  `alpha` varchar(16) default NULL,
```

```
  `src_begin_date` datetime default NULL,
  `bravo` varchar(16) default NULL,
  `grid` varchar(8) default NULL,
  `east_grid_ref` int(11) default NULL,
  `north_grid_ref` int(11) default NULL,
  `charlie` varchar(16) default NULL,
  `post_code` varchar(16) default NULL,
  `src_end_date` datetime default NULL,
  `elevation` smallint(6) default NULL,
  `delta` varchar(16) default NULL,
  `echo` varchar(16) default NULL,
  `foxtrot` varchar(16) default NULL,
  `golf` varchar(16) default NULL,
  `hotel` varchar(16) default NULL,
  `india` varchar(16) default NULL,
  `juliet` varchar(16) default NULL,
  PRIMARY KEY  (`src_id`)
)
```

A.1.2 Data grouping – first attempt

The first attempt of selecting the stations that are supposed to be active was done with the following SQL query which yielded a list of 9071 stations:

```
SELECT s.src_id, s.src_name, s.src_begin_date,
s.src_end_date, s.lon, s.lat FROM stations s
WHERE s.src_begin_date <= '2001-01-01' AND
s.src_end_date >= '2007-12-31'
```

This list should now be classified by regions. However, the regions must first be defined.

A.1.3 Sorting by regions

Before sorting, the regions were described with geographical coordinates. Google Earth was used for this. Sixteen polygons were overlaid on Great Britain. The data of the polygons was exported into a KMZ file which is a zip-compressed version of a KML file. KML stands for 'Keyhole Markup Language' and it "uses an XML structure with nested elements and attributes" to display geographic data [10].

Preparing the sources

The coordinates for each point of the polygon representing the region were extracted by hand and edited on jEdit to organise them as a simple table rather than a long string. The transformation was done with a macro written in the Beanshell scripting language.

```
void sliceLine()
{
    line = textArea.getCaretLine();
    text = textArea.getLineText(line);
    lines = text.split(" ", 0);

    textArea.setSelectedText("lon,lat\n");
    for(int i=0;i<lines.length-1;i++)
    {
        coords=lines[i].split(",", 0);
        textArea.setSelectedText(coords[0]+","+coords[1] + "\n");
    }
    textArea.deleteLine();
}
sliceLine();
```

The source string is something like this (cropped so it can fit the page):

```
-5.136081265489793,49.88449959913417,0 -3.693415209843323, ...
```

The polygon in the KML file is defined as a sequence of points, where the first and the last are the same. Note that the order of data is longitude, latitude and elevation. The output we want is something like the following (omitting the last point and using only longitude and latitude):

```
lon,lat
0.3159496200555481,51.47551155828858
1.004590983427008,51.51028161869819
1.844850791349728,52.23038416140323
1.873694346066315,52.9258543388793
0.3960541288590775,53.06898974157086
-0.0775780683329117,52.36496478085275
-0.8480843959788729,51.92583653631429
-0.5442039526437981,51.6759853251772
0.1496691693183436,51.67073087653791
```

The output is saved as a text file and the process is repeated for each region.

Loading regions into R

The following code loads each file defining a region and stores it in the R object `romeo`.

```
romeo<-list()
romeo[[1]]<-read.csv('src/northernIlesRegion.csv',header=T,row.names=NULL)
romeo{[2]]<-read.csv('src/northWestScotlandRegion.csv',header=T,row.names=NULL)
romeo[[3]]<-read.csv('src/northEastScotlandRegion.csv',header=T,row.names=NULL)
romeo[[4]]<-read.csv('src/southScotlandRegion.csv',header=T,row.names=NULL)
romeo[[5]]<-read.csv('src/northWestEnglandRegion.csv',header=T,row.names=NULL)
romeo[[6]]<-read.csv('src/northEastEnglandRegion.csv',header=T,row.names=NULL)
romeo[[7]]<-read.csv('src/eastEnglandRegion.csv',header=T,row.names=NULL)
romeo[[8]]<-read.csv('src/westMidlandsRegion.csv',header=T,row.names=NULL)
romeo[[9]]<-read.csv('src/eastMidlandsRegion.csv',header=T,row.names=NULL)
romeo[[10]]<-read.csv('src/eastAngliaRegion.csv',header=T,row.names=NULL)
romeo[[11]]<-read.csv('src/northWalesRegion.csv',header=T,row.names=NULL)
romeo[[12]]<-read.csv('src/southWalesRegion.csv',header=T,row.names=NULL)
romeo[[13]]<-read.csv('src/londonRegion.csv',header=T,row.names=NULL)
romeo[[14]]<-read.csv('src/southWestEnglandRegion.csv',header=T,row.names=NULL)
romeo[[15]]<-read.csv('src/southEnglandRegion.csv',header=T,row.names=NULL)
romeo[[16]]<-read.csv('src/southEastEnglandRegion.csv',header=T,row.names=NULL)

regionNames<-c('ni','nws','nes','ss','nwe','nee','ee',
'wm','em','ea','nw','sw','ldn','swe','se','see');
names(romeo)<-regionNames
```

Loading stations into R

The next part is done by sourcing the station list from MySQL into R. The following code loads the necessary libraries, connects R with MySQL, executes the SQL query and fetches the result into `sierra`.

```
library(RMySQL);
mysql <- MySQL();
cxn <- dbConnect(mysql, user='root', password='',
dbname="whisky", host="localhost");

sql<-"SELECT s.src_id, s.src_name, s.lon, s.lat
```

```
FROM stations s
WHERE s.src_begin_date <= '2001-01-01' AND
s.src_end_date >= '2007-12-31'"

query <- dbSendQuery(cxn, sql)
sierra <- fetch(query, n=-1)
```

Determine if a point lies within a polygon

The core of sorting the stations into regions is to know whether a point is inside a polygon. The code for this algorithm is based on one provided in [5]. It was adapted from the C programming language to R.

```
polygonContainsPoint<-function(point, region)
{
    counter <- 0
    n <- length(region[ ,1])
    index <- 1:n
    index <- rep(index, 2)
    index <- index[seq(2, length=n)]

    p1 = region[1,]

    for(i in 1:n)
    {
        p2 = region[index[i], ]
        if (point[2] > min(p1[2], p2[2]))
        {
            if (point[2] <= max(p1[2], p2[2]))
            {
                if (point[1] <= max(p1[1], p2[1]))
                {
                    if (p1[2] != p2[2])
                    {
                        xinters = (point[2]-p1[2]) * (p2[1]-p1[1]) /
                                    (p2[2] - p1[2]) + p1[1]
                        if (p1[1] == p2[1] || point[1] <= xinters)
                            counter <- counter + 1
                    }
                }
            }
```

```
        }
        p1 <- p2
    }

    if (counter %% 2 == 0)
        F
    else
        T
}
```

Sorting

We now have all the elements to construct a script that sorts the stations by regions. The R script is the following:

```
n<-length(romeo);
m<-length(sierra[,1]);
theChosen<-c();

for(i in 1:m)
{
    for(j in 1:n)
    {
        if(polygonContainsPoint(unlist(sierra[i,3:4]), romeo[[j]]))
        {
            theChosen<-append(theChosen,
                              c(regionNames[j], unlist(sierra[i,])) );
            break;
        }
    }
}
```

To finish, we rearange the result from a vector to a matrix, sort it according to the column that contains the region code (e.g. ea, ee, nes, etc.) and write the matrix to a file for future use.

```
charlie<-matrix(theChosen, ncol=6, byrow=T)
sortindex<-sort(charlie[,1],index.return=T)
charlie<-charlie[sortindex[[2]],]
write.matrix(charlie,'output/worthyStations.txt')
```

But as it was later discovered, this set of stations was insufficient. Instead, we seek the stations that actually recorded observations. Preferably,

those that recorded reliable observations. So, before selecting the stations, we filter the observation data and then find out which stations recorded it.

A.2 Filtering process

The following shows how the filtering process was done.

A.2.1 Observation table definition

The wind speed observation data was loaded to a table named 'wind_data' with the following structure:

```
CREATE TABLE  `whisky`.`wind_data` (
  `ob_end_time` datetime NOT NULL,
  `id_type` varchar(16) NOT NULL,
  `id` varchar(32) NOT NULL default '0',
  `ob_hour_count` int(11) NOT NULL,
  `met_domain_name` varchar(32) default NULL,
  `version_num` int(11) NOT NULL,
  `src_id` int(11) default NULL,
  `rec_st_ind` varchar(16) default NULL',
  `mean_wind_dir` varchar(16) default NULL,
  `mean_wind_speed` varchar(16) default NULL,
  `wind_speed` double default NULL,
  `max_gust_dir` varchar(16) default NULL,
  `max_gust_speed` varchar(16) default NULL,
  `max_gust_ctime` varchar(16) default NULL,
  `mean_wind_dir_q` varchar(16) default NULL,
  `mean_wind_speed_q` varchar(16) default NULL,
  `max_gust_dir_q` varchar(16) default NULL,
  `max_gust_speed_q` varchar(32) default NULL,
  `max_gust_ctime_q` varchar(32) default NULL,
  `meto_stmp_time` varchar(32) default NULL,
  `midas_stmp_etime` varchar(32) default NULL,
  `mean_wind_dir_j` varchar(32) default NULL,
  `mean_wind_speed_j` varchar(32) default NULL,
  `max_gust_dir_j` varchar(32) default NULL,
  `max_gust_speed_j` varchar(32) default NULL,
  KEY `ob_end_time` (`ob_end_time`),
  KEY `id_type` (`id_type`),
  KEY `id` (`id`),
```

```
  KEY `met_domain_name` (`met_domain_name`),
  KEY `src_id` (`src_id`),
  KEY `rec_st_ind` (`rec_st_ind`),
  KEY `mean_wind_speed_q` (`mean_wind_speed_q`)
)
```

The MIDAS Land Surface Stations data for wind that was used is the one contained in the following files. They can be downloaded from (prior registration) http://badc.nerc.ac.uk/browse/badc/ukmo-midas/data/WM/yearly_files.

- midas_wind_200101-200112.txt
- midas_wind_200201-200212.txt
- midas_wind_200301-200312.txt
- midas_wind_200401-200412.txt
- midas_wind_200501-200512.txt
- midas_wind_200601-200612.txt
- midas_wind_200701-200712.txt

The next script was used to load one of the 'comma separated value' files (CSV) from MIDAS. The other files were loaded similarly:

```
LOAD DATA LOCAL INFILE
'/home/pablo/MSc/sandbox/data/midas_wind_200101-200112.txt'
REPLACE INTO TABLE whisky.wind_data
FIELDS TERMINATED BY ','
ENCLOSED BY ''
LINES TERMINATED BY '\n'
IGNORE 1 LINES;
```

The data is now in place to manipulate it.

A.2.2 Stations that recorded data

A first glimpse of which stations actually recorded data can be achieved by the following SQL query: As it was noted at the end of section A.1, the first use of filtering the data is to select the stations that recorded it. This is achieved by the following SQL query:

Status code	Version	Record status indicator	Count
1001	0	Normal ingestion of observation	146
1001	1	at creation	1,087,825
1004	1	Receive a correction before normal observation received	2,567
1010	0	Start of QC; observation	499
1010	1	extracted for QC checks	3,703
1011	0	QC level raised on Version 1	23,341
1011	1		11,219,169
1012	0	Create Version 0 First QC amend	159,216
1012	1	to an attribute other than just change of QC level	153
1013	0	Version 0 exists with no version 1	39,048
1022	0	Version 1 Creation, Version 0 is	11
1022	1	frozen as original data state indicator 1012	59,726
		Total:	12,595,404

Table A.1: The different record states for each version and their frequency. Source of status descriptions [24]

```
SELECT d.src_id, s.src_name, COUNT(d.src_id)
FROM wind_data d INNER JOIN stations s ON d.src_id=s.src_id
GROUP BY d.src_id ORDER BY count(d.src_id);
```

A.2.3 Record status indicator

Table A.1 was obtained with the following SQL query:

```
SELECT rec_st_ind, version_num, COUNT(rec_st_ind)
FROM wind_data
GROUP BY rec_st_ind, version_num
ORDER BY COUNT(rec_st_ind);
```

A.2.4 q value

Figure A.1 shows a table diagram and was obtained with the following SQL query:

```
SELECT met_domain_name, COUNT(met_domain_name) FROM wind_data w
GROUP BY met_domain_name
ORDER BY COUNT(met_domain_name) DESC;
```

	4th Digit	3rd Digit	2nd Digit	1st Digit	
CODE	4 Indicates 1 of up to 8 remarks about an estimate	3 Indicates 1 of up to 8 possible descriptions of the value of the element	2 Indicates 1 of up to 8 statements about the original value	1 Indicates which of 10 possible stages of QC has been reached	COUNT
0					1,207,794
1000	Estimate/correction derived automatically from a program with no manual intervention	Observed and not suspect	Original value is/was not queried, or no information available	Initial climate QC program not run	147,85
120		Observed and suspect (i.e. has failed the latest QC check), or there are strong grounds for suspecting the accuracy of the observation	Failed MIDAS validation		3
1					8,016,812
301		An estimate where the original value is missing and cannot be retrieved	Original value is/was not queried, or no information available	Initial climate QC program has run	2
2001	Estimate/corrected value has been set manually (with or without assistance from a program)	Observed and not suspect			12,011
2301		An estimate where the original value is missing and cannot be retrieved			3,706
141		Observed and suspect (i.e. has failed the latest QC check), or there are strong grounds for suspecting the accuracy of the observation	Failed climate QC range check		2,227
151			Failed climate QC internal consistency check		15,825
2151	Estimate/corrected value has been set manually (with or without assistance from a program)				148
161			Failed climate QC sequence check		7,345
2161	Estimate/corrected value has been set manually (with or without assistance from a program)				3
6					2,758,068
2006	Estimate/corrected value has been set manually (with or without assistance from a program)	Observed and not suspect	Original value is/was not queried, or no information available	Final (or only) areal or buddy job run and queries processed	965
No flag					555,709

Figure A.1: The q flag for the wind speed and its occurrence frequency

Network	Count
HCM	9,319,867
AWSHRLY	2,340,033
HWND6910	567,099
SYNOP	291,958
HWNDAUTO	57,217
DLY3208	19,230
Total:	12,595,404

Table A.2: The different networks and their occurrence frequency

A.2.5 Value precision

The following SQL query counts how many records have integer values as their reported wind speed. The result of the query is 12,207,145.

```
SELECT COUNT(mean_wind_speed)
FROM wind_data d WHERE mean_wind_speed REGEXP '^[0-9]+$';
```

A.2.6 Observation domain

Table A.2 was obtained with the following SQL query:

```
SELECT met_domain_name, COUNT(met_domain_name) FROM wind_data w
GROUP BY met_domain_name
ORDER BY COUNT(met_domain_name) DESC;
```

A.2.7 Filter query

As explained at the end of section 2.3.4 there are four filters to create:

- Belong to the HCM domain
- Have 1011 as status indicator and 1 as version
- Have an integer value as mean wind speed
- Have 1 or 6 flags as q value

As it was noted at the end of section A.1, the first use of filtering the data is to select the stations that recorded it. This is achieved by the following SQL query:

```
SELECT d.src_id, s.src_name, COUNT(d.src_id)
FROM wind_data d INNER JOIN stations s ON d.src_id=s.src_id
WHERE d.met_domain_name REGEXP 'HCM'
AND d.rec_st_ind REGEXP '1011'
AND version_num=1
AND d.mean_wind_speed_q REGEXP '1|6'
AND d.mean_wind_speed REGEXP '^[0-9]+$'
GROUP BY d.src_id ORDER BY count(d.src_id);
```

Isolating the filtered values

SQL queries such as the one above are time consuming to execute. It is impractical to repeatedly execute the filtering query each time data should be fetched for analysis. To avoid this, a new table is created based only on the records returned by the filtering query. This has the advantage of not repeating the query and of having a smaller lookup index which speeds-up future queries. This is done as follows:

```
CREATE TABLE `whisky`.`power` (`wind_speed` double default NULL)
SELECT ob_end_time, id_type, id, ob_hour_count,
met_domain, version_num, src_id, rec_st_ind,
mean_wind_speed WHERE d.met_domain_name REGEXP 'HCM'
AND d.rec_st_ind REGEXP '1011'
AND version_num=1
AND d.mean_wind_speed_q REGEXP '1|6'
AND d.mean_wind_speed REGEXP '^[0-9]+$'
```

A.2.8 Unit and height conversion

A new column 'wind_speed' is created to store the new value converted to $^{m}/_{s}$ and extrapolated to 60 m. The column is created empty, but with the following update query the value is set:

```
UPDATE power
SET wind_speed = CONVERT(mean_wind_speed, DECIMAL) * (0.6645064)
```

A.2.9 Data grouping – second attempt

In section ?? above, a list of stations that recorded reliable results is obtained. The process described in section A.1.3 is repeated but with the records from the new table "power".

Code	Count
ea	5
ee	5
em	11
ldn	1
nee	8
nes	11
ni	3
nw	8
nwe	7
nws	7
se	14
see	6
ss	8
sw	6
swe	11
wm	5

Table A.3: The number of stations for each region

The final result of the sorting process is the following list of stations classified by region:

A.2.10 Demand data

The demand data obtained from National Grid was also loaded into the database. Below is the table definition to contain the data:

```
CREATE TABLE ‘demand‘ (
  ‘datetime‘ datetime NOT NULL,
  ‘indo‘ int(10) unsigned NOT NULL,
  ‘gb‘ int(10) unsigned NOT NULL,
  ‘dem‘ int(10) unsigned NOT NULL,
  ‘tgsd‘ int(10) unsigned NOT NULL,
  PRIMARY KEY  (‘datetime‘),
  KEY ‘tgsd‘ (‘tgsd‘)
)
```

The SQL query to select the maximum value from each hour period is:

```
SELECT datetime, max(tgsd) as demand FROM demand
WHERE datetime BETWEEN ’2001-06-01 00:00:00’ AND ’2002-07-01 23:59:59’
```

```
GROUP BY year(datetime), month(datetime), day(datetime), hour(datetime)
ORDER BY datetime
```

A.3 Information extraction

The final part of the implementation of the methodology is to extract and plot the information that is sought.

A.3.1 Power curve function

To fit an equation to the power curve for the Bonus 2 MW the following code was used. First the points data is loaded, next the curve is fitted, and finally, the function is modified so that it includes the cut-in and cut-out speeds (see figure 2.2; the circles are the points given and the solid line is the resulting function).

```
bonus<-read.table('sources/bonus-powercurve.txt',header=T,row.names=NULL)
powerfun<-splinefun(bonus)

powerfoo<-function (x)
{
    x[x<4]<-0
    x[x>25]<-0
    y<-powerfun(x)
    y[y<powerfun(4)]<-0
    y
}
```

A.3.2 Weights data

The data for the weights was obtained from the BWEA webpage. The BWEA provides four listings of the operational, in construction, consented and in planning wind farms. The listings contain the wind farm MW capacities and their coordinates, among other things.

It is necessary to classify the wind farms into regions and then sum the capacity for each one to obtain its weight. The data from the listings needs to be extracted from an HTML page to a clean text file. The following steps were taken to automate the data extraction.

First, the HTML code that defined the table in the webpage was copied and pasted in a new file. This file was later imported into Excel. The coordinate data in the table is in the Degree Minute Second (DMS) form and must be changed to a decimal representation. So an Excel macro was used. The original code is from [16] and was modified to suit this particular dataset.

```
Function Convert2Decimal(coord As String) As Double
   ' Declare the variables to be double precision floating-point.
   Dim degrees As Double
   Dim minutes As Double
   Dim seconds As Double
   Dim dir As String
   Dim returnValue As Double

   Dim sploof() As String

   sploof = Split(Trim$(Replace$(coord, Chr(160), vbNullString)), " ")

   degrees = Val(sploof(0))
   minutes = Val(sploof(1)) / 60
   seconds = Val(StrReverse(Mid(StrReverse(sploof(2)), 1))) / 3600

   returnValue = degrees + minutes + seconds

   dir = Trim(Right(sploof(2), 1))
   If (StrComp(dir, "W") = 0 Or StrComp(dir, "S") = 0) Then
        returnValue = returnValue * -1
   End If

   Convert2Decimal = returnValue

End Function
```

Once the coordinates were converted from DMS to decimal, the four datasets were grouped into one and exported to a CSV file. This file was fed to the script in section A.1 and the output was table 2.2.

A.3.3 Aggregate & Merge

The calculation defined in section 2.3.6 is done in R with the `aggregate` function. Once all the observations made by the stations of a region have been fetched, the hourly average is calculated. Next, the averaged wind speed is converted into hourly capacity factor. To finish, the matching with peak hours (section 2.4.2) is done with `merge`. The code follows:

```
query <- dbSendQuery(cxn, sql)
cf<-fetch(query, n=-1)
cf[,2]<-powerfoo(cf[,2]) #calculate power output
cf[,2]<-(cf[,2]/2000)*100 #make CF (normalise)
cf<-aggregate(cf[,2], list(cf[,1]), mean)

peakcf<-merge(cf,peakhours,all=F)
```

This code results in a dataset of hourly capacity factors that match the hours identified as peak demand (whether it be 5%, 10%, etc).