John Field Consulting Pty Ltd

ABN 62 096 615 235

10 High St, Burnside SA 5066
Telephone: (08) 8332 5294
Facsimile: (08) 8332 1229
Mobile: 0409 097 586
Email: JohnField@ozemail.com.au

Distribution of SAIDI data

Report produced for

United Energy
Level 3, 501 Blackburn Road
MOUNT WAVERLEY Vic. 3149

Version 2, 26 October 2009

Distribution of SAIDI data

EXECUTIVE SUMMARY

The Australian Energy Regulator's STPIS assumes that the logarithm of SAIDI is normally distributed, or can be made normally distributed by an appropriate transformation. This report examines the distribution of United Energy's SAIDI data. The data used comprised daily unplanned SAIDI for the last five calendar and financial years.

There does not appear to be much difference between calendar years and financial years.

The five years of log(SAIDI) values are not normally distributed. The distribution has a small positive skewness (ie the upper tail is longer than the lower tail) and the distribution is rather more 'flattened' than a normal distribution, with more bulk in the centre and less in the tails than we might expect in a normal distribution.

A Box-Cox transformation does remove the skewness, but not the 'flatness' of the distribution, and the transformed distribution cannot be regarded as normally distributed. No better transformations were found.

An examination of the variability in log(SAIDI) was carried out. Within the one season and the one year, there is evidence that the distribution is approximately normal, but the mixture of distributions over seasons and years results in a non-normal distribution. There is greater evidence of stability of log(SAIDI) values in summer and spring than winter and autumn.

In summary, it does not appear that a better transformation for SAIDI than log(SAIDI) can be found.

TABLE OF CONTENTS

1. Introduction and data

The Australian Energy Regulator (AER) has proposed a methodology to identify Major Event Days from a record of daily SAIDI values. This methodology assumes that logarithm of SAIDI is normally distributed, or can be made normally distributed by an appropriate transformation[1].

This report examines the distribution of United Energy's SAIDI data.

Data supplied by United Energy comprised daily unplanned SAIDI data for the period 1/01/2004 to 31/08/2009, excluding load shedding and transmission line failures.

During this period of 2070 days, SAIDI ranged from 0.000091 to 232.28 minutes. However, only 10 days had values in excess of 5 minutes.

Calculations and graphics in this report were carried out using the R statistical language[2] except where otherwise noted. A summary of the R code used is given in the Appendix.

[1] AER (2009) Electricity distribution network service providers – Service target performance incentive scheme, version 1.2, September 2009.

[2] R Development Core Team (2009). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

2. Distribution of log(SAIDI)

Since the Australian Energy Regulator (AER) considers data in five year periods, we consider the last five years' data for both calendar and financial years.

2.1 Calendar years

A histogram of log(SAIDI) over the 5-year period 2004-2008 is plotted in Figure 1. (We use log throughout to refer to logarithms to the base e).

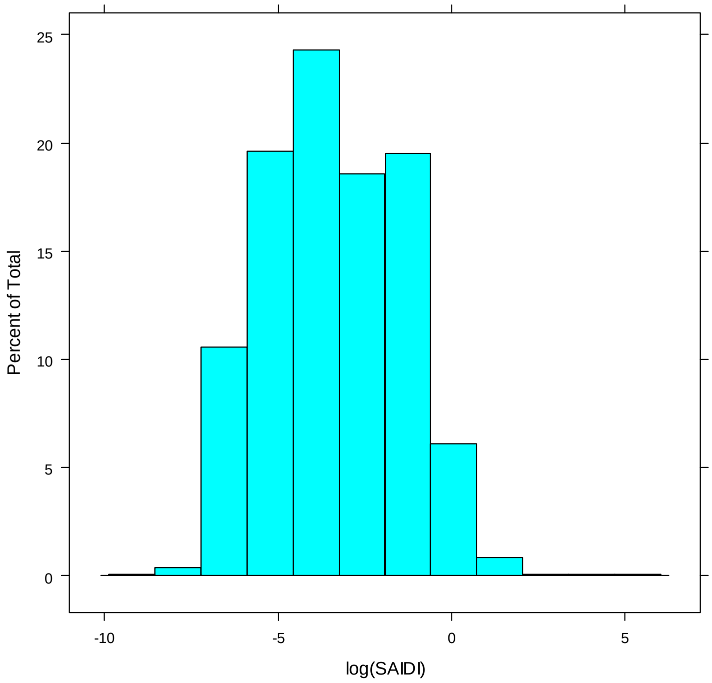

Figure 1: Histogram of log(SAIDI), 2004 – 2008

Basic descriptive data about the distribution of log(SAIDI) is shown in Table 1.

Table 1: Summary statistics for log(SAIDI), 2004 – 2008

Mean	-3.41
Median	-3.52
Skewness	0.15 (95% ci = 0.04 – 0.26)
Kurtosis	-0.53 (95% ci = -0.75 – -0.30)

For a normal distribution, both the skewness and kurtosis will equal zero. The confidence intervals show that both are significantly different to zero, since zero is not included in the confidence interval. The distribution is slightly skewed to the right (ie has a longer upper tail than lower) which is confirmed by a slightly larger mean than median. The distribution is also platykurtic, ie is rather more ‘flattened’ than a normal distribution, with more bulk in the centre and less in the tails than we might expect in a normal distribution.

The corresponding normal probability plot is shown in Figure 2. Quantiles of the normal distribution are plotted on the x axis and corresponding quantiles of the data are on the y axis. If log(SAIDI) is normally distributed, we would expect the points to approximate a straight line, and to lie mainly between the 95% pointwise confidence limits shown on the plot. Departures from this straight line or points outside the confidence limits indicate departures from normality. The straight line is a robust regression line.

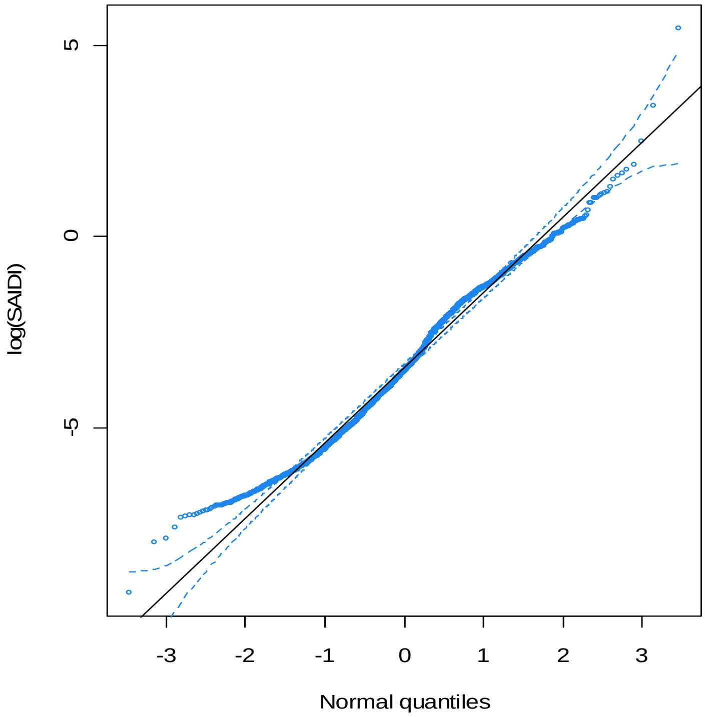

Figure 2: Normal probability plot for log(SAIDI) 2004 – 2008

There are clear departures from a normal distribution, particularly in the lower tail, but also in the upper tail and the middle of the distribution. These departures are consistent with a mixture of normal distributions and/or a light-tailed distribution compared to a normal distribution.

Formally, we can use various tests for normality. Results from a variety of tests are shown in Table 2.

Table 2: P values for normality tests, log(SAIDI) 2004-2008

Anderson-Darling	$P<2x10^{-16}$
Cramer-von Mises	$P=4x10^{-10}$
Lilliefors	$P=8x10^{-10}$
Shapiro-Wilk	$P=9x10^{-13}$

These confirm the non-normality of the data.

2.2 Financial years

The available five financial years run from 1/07/04 to 30/06/09. A histogram for this period is shown in Figure 3.

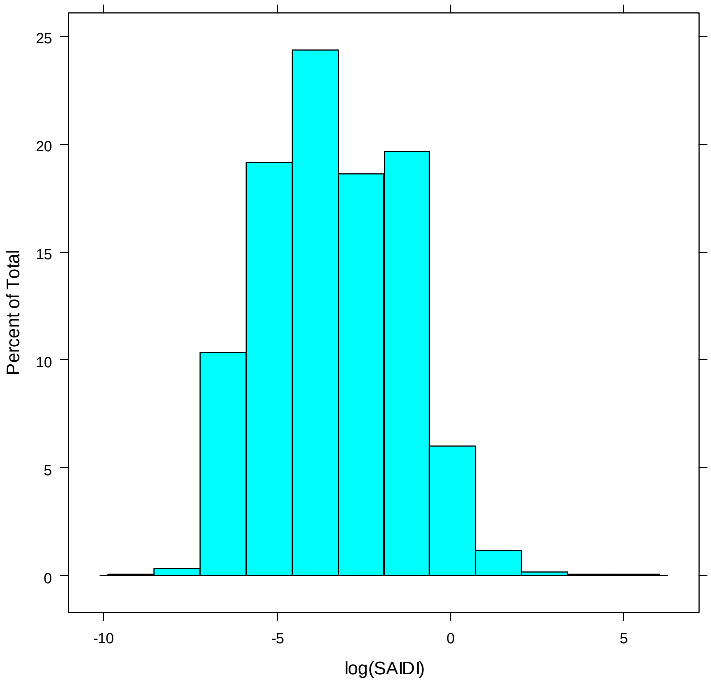

Figure 3: Histogram of log(SAIDI), 04/05 – 08/09

The normal probability plot is shown in Figure 4. The behaviour of log(SAIDI) is obviously similar over both calendar and financial years.

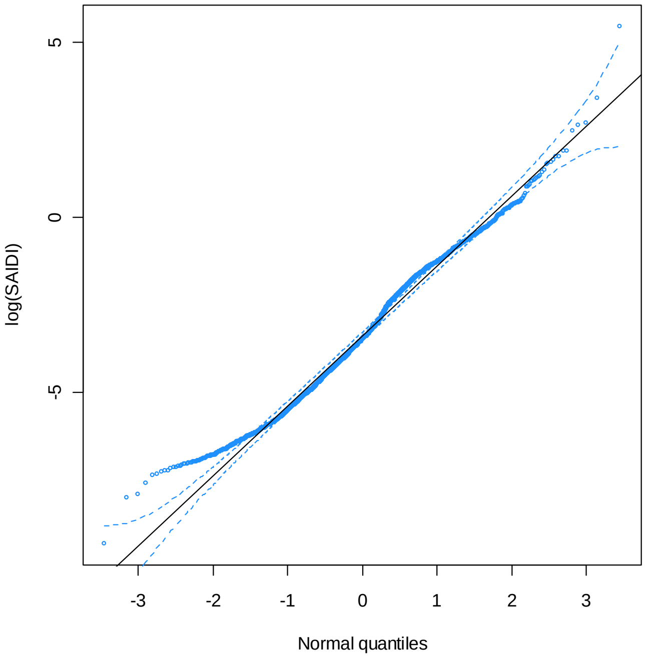

Figure 4: Normal probability plot for log(SAIDI), 04/05 - 08/09

Basic descriptive data about the distribution of log(SAIDI) is shown in Table 3.

Table 3: Summary statistics for log(SAIDI) 04/05 - 08/09

Mean	-3.37
Median	-3.49
Skewness	0.18 (95% ci = 0.07 – 0.30)
Kurtosis	-0.45 (95% ci = -0.68 – -0.23)

The results are very similar to those for the calendar years.

Formal significance tests for normality are shown in Table 4.

Table 4: P values for normality tests, log(SAIDI), 04/05 - 08/09

Anderson-Darling	$P<2x10^{-16}$
Cramer-von Mises	$P=5x10^{-10}$
Lilliefors	$P=1x10^{-7}$
Shapiro-Wilk	$P=4x10^{-12}$

2.3 Comments

There does not appear to be much difference between calendar years and financial years.

The five years of log(SAIDI) values are not normally distributed. The formal tests of significance show this, but with such a large data set (1826 or 1827 days), almost any formal test will show a significant effect. Additionally, the tests assume independent data samples, and strictly, that is not so with these data – there is a small but significant serial correlation between values, driven largely by the weather, since small SAIDI values tend to be followed small values, and large ones tend to be followed by large ones.

The most useful assessment of normality can be made from the normal probability plots. Because of the large number of data points, the confidence limits shown on the plot are very close together. However the systematic discrepancies suggest either (or both) a light-tailed distribution and/or a mixture of normal distributions. We follow this further in Section 4 below.

3. Box-Cox transformation

The Box-Cox transformation[3] for a variable X is defined as

$$X^{(\lambda)} = (X^{\lambda} - 1) / \lambda \text{ for } \lambda \neq 0, \text{ and } X^{(\lambda)} = \log(X) \text{ for } \lambda = 0.$$

The parameter λ can be fitted by maximum likelihood, in the current case using the 'car' package[4] of the R statistical language.

We look at the transformation separately for calendar and financial years.

3.1 Calendar years

The Box-Cox parameter, λ, for this data set is -0.0331. This is significantly different from zero (equivalent to a log transform), with P=0.003.

The histogram and normal probability plot for the Box-Cox transformed data is shown below.

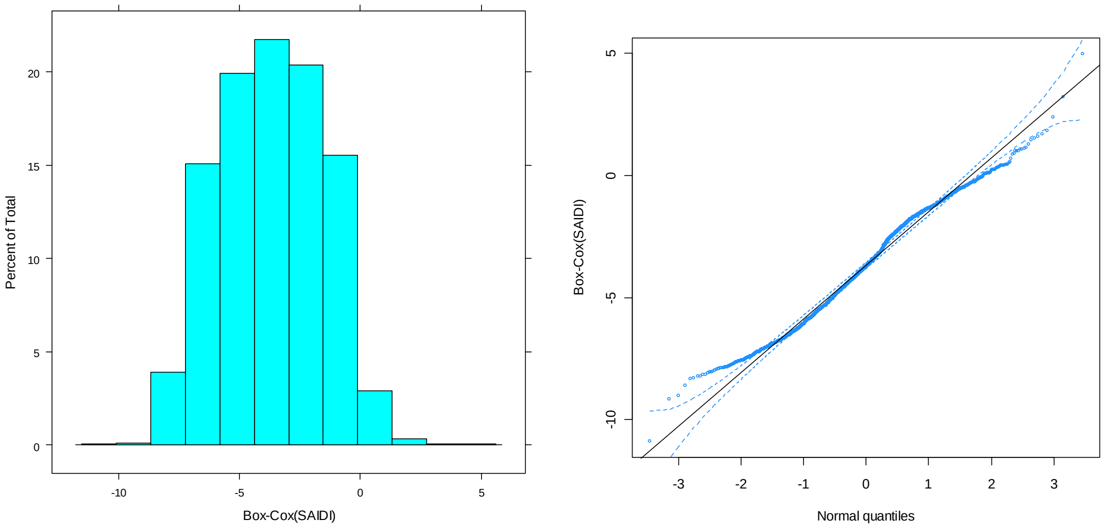

Figure 5: Histogram and normal plot of Box-Cox(SAIDI) 2004 – 2008

The same patterns are seen in the probability plot as in the log transformed SAIDI data.

Summary statistics are shown in Table 5 and test results in Table 6.

[3] Box, GEP and Cox DR (1964) An analysis of transformations (with discussion). *Journal of the Royal Statistical Society B* **26**, pp 211-252.

[4] John Fox (2009). car: Companion to Applied Regression. R package version 1.2-14. http://CRAN.R-project.org/package=car

Table 5: Summary statistics for Box-Cox transformed data, 2004-2008

Parameter	Estimate
Mean	-3.68
Median	-3.74
Skewness	0.01 (-0.10 – 0.12)
Kurtosis	-0.65 (-0.88 - -0.43)

The transformation has reduced the skewness – it is not significantly different from zero – note that the confidence interval contains the zero value.

Table 6: P values for normality tests, Box-Cox transformed data, 2004-2008

Test	Result
Anderson-Darling	$P<2x10^{-16}$
Cramer-von Mises	$P=4x10^{-10}$
Lilliefors	$P=2x10^{-12}$
Shapiro-Wilk	$P=4x10^{-12}$

3.2 Financial years

The Box-Cox parameter is -0.0398, significantly different to zero (P=0.0003).

The histogram and normal probability plots are shown below.

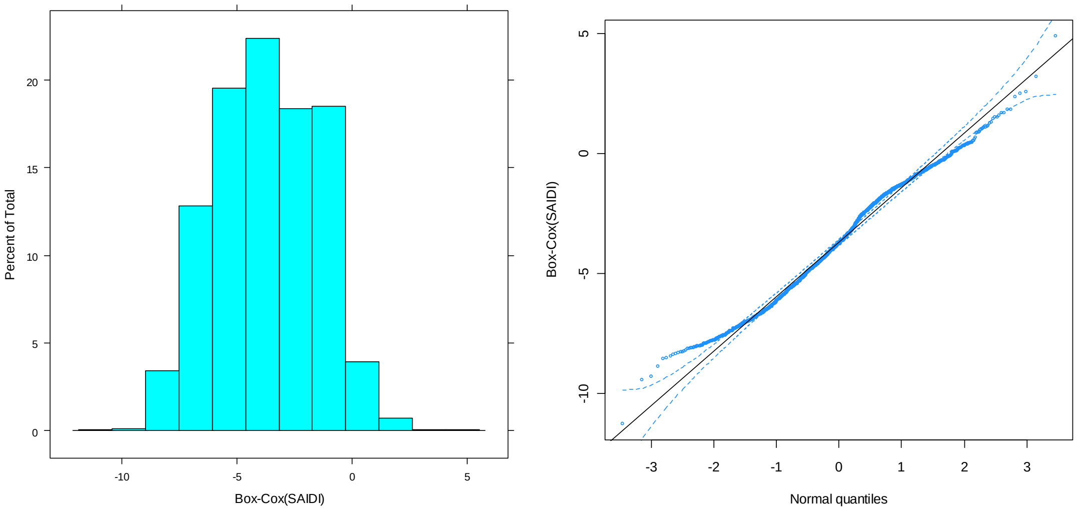

Figure 6: Histogram and normal plot of Box-Cox(SAIDI), 03/04 – 08/09

Distributional parameters are shown in Table 7 and test results in Table 8.

Table 7: Summary statistics for Box-Cox transformed data, 03/04 - 08/09

Parameter	Estimate
Mean	-3.69
Median	-3.74
Skewness	0.02 (-0.10 – 0.13)
Kurtosis	-0.62 (-0.84 – -0.39)

Again the skewness has been removed, but not the kurtosis. The same abnormalities appear in the normal probability plot.

Table 8: P values for normality tests, Box-Cox transformed data, 03/04 - 08/09

Test	Result
Anderson-Darling	$P<2\times10^{-16}$
Cramer-von Mises	$P=6\times10^{-10}$
Lilliefors	$P=8\times10^{-12}$
Shapiro-Wilk	$P=5\times10^{-11}$

3.3 Comments

The Box-Cox transformation has made a minor improvement to the distributions in reducing the skewness of the data, but the transformed data are still clearly non-normal with the same problems as log(SAIDI). The formal tests of significance still give significant results with extremely small P values. The Box-Cox transformation simply does not normalise these data, and working with log(SAIDI) will be simpler and provide almost identical results.

Minitab has a function to test the fit of a range of distributional transformations, although some of these are clearly not appropriate for these data. None of these (either the two or three-parameter versions where applicable) fitted the data: lognormal, lognormal, exponential, exponential, Weibull, Extreme Value, Gamma, Logistic, Loglogistic.

4. Yearly variation

To further examine the distributional variability, and its departure from normality, we plot below probability plots for each calendar year (Figure 7) and financial year (Figure 8). There are five full years represented on each plot.

The general shape is the same for each year, and there is a rough ordering of years, with the earliest year at the bottom of the band of lines and the latest year at the top. This ordering is not exact throughout, but it suggests movement in the distribution with changing years.

However the shape is fairly consistent over years, whether calendar or financial. Hence we use all data for the remainder of this section.

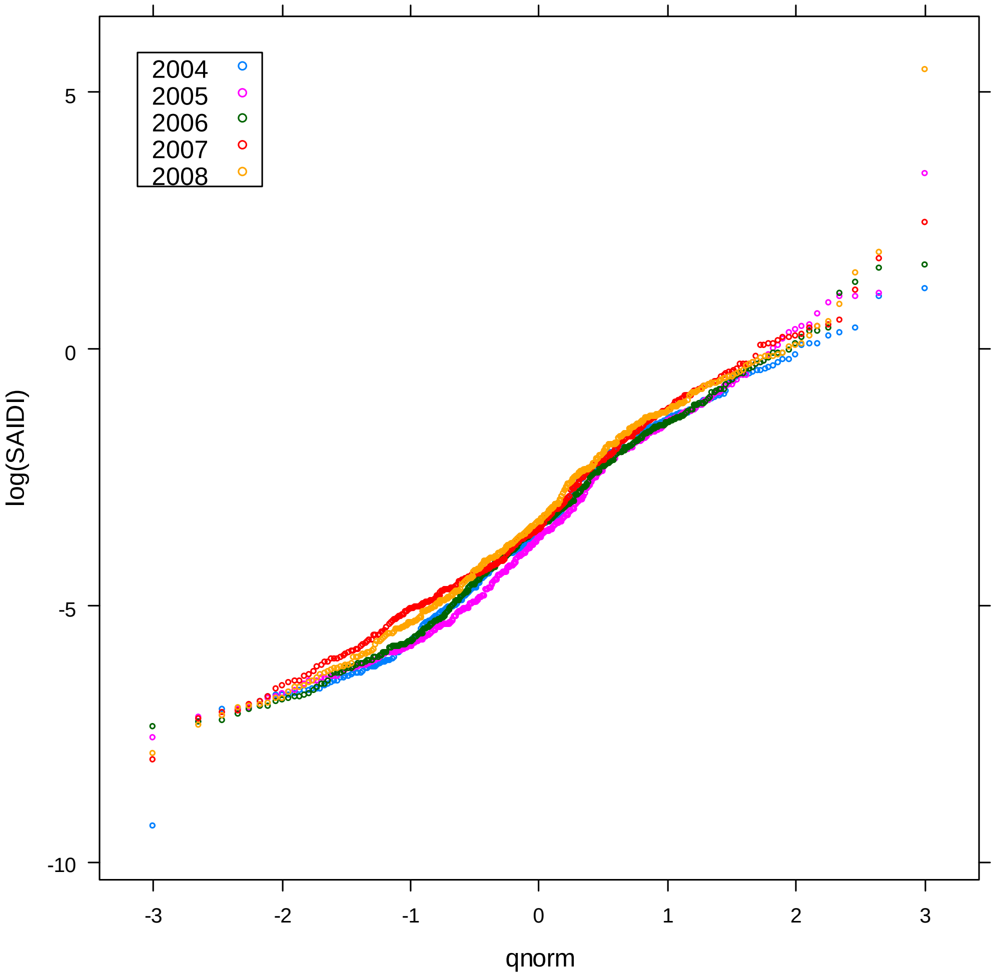

Figure 7: Normal probability plots for calendar years

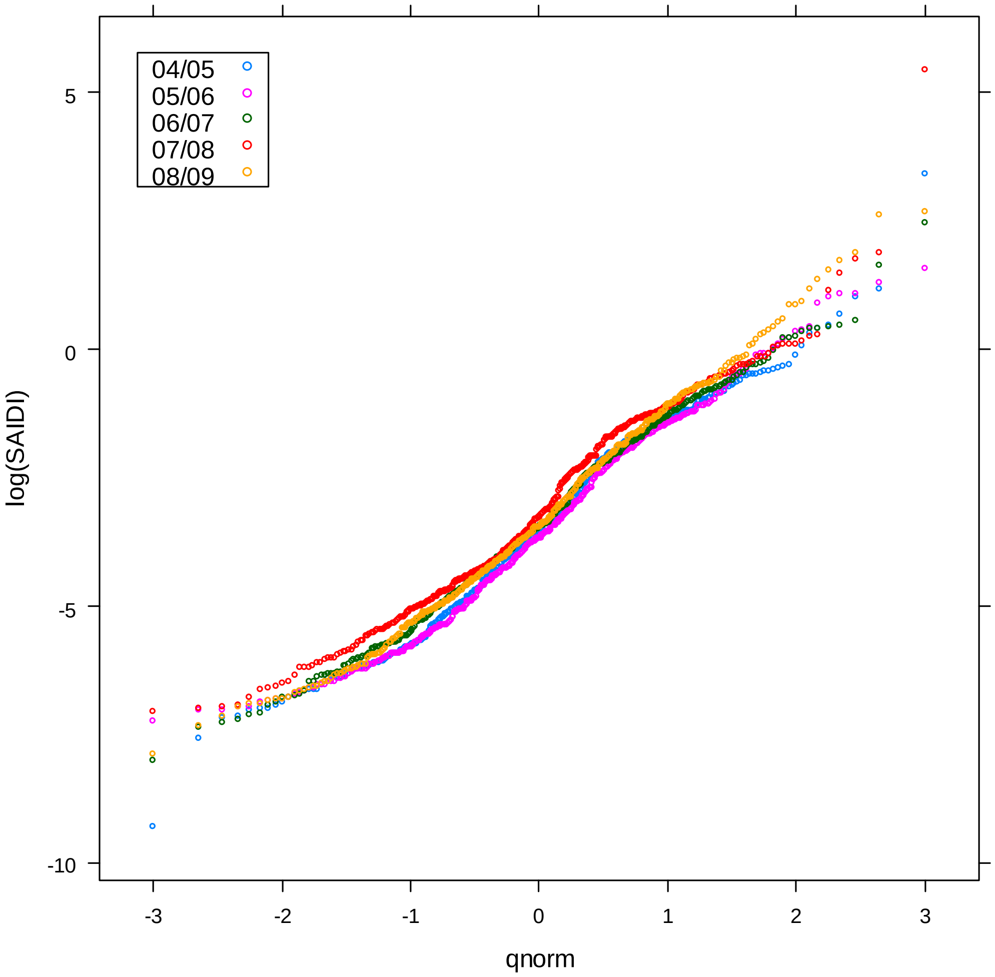

Figure 8: Normal probability plots for financial years

The gradation in years suggests that the log(SAIDI) values are increasing with year. To examine this idea we split the year into approximate seasons:

Summer	Dec – Mar
Autumn	Apr – May
Winter	Jun – Aug
Spring	Sep – Nov

Note that not all 'seasons' are the same length. To avoid splitting seasons over years, we regard a 'year' as the period December to November.

We use boxplots[5] to examine the distribution of log(SAIDI) in each season (Figure 9).

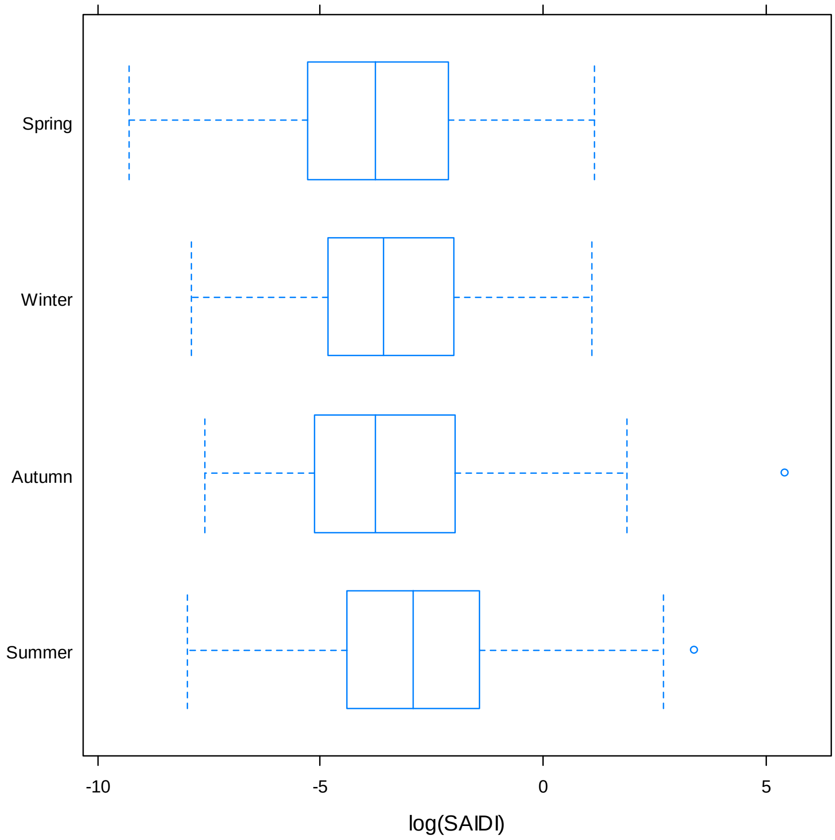

Figure 9: Seasonal variation in the distribution of log(SAIDI)

The data include all years. The median log(SAIDI) (and hence SAIDI) is greatest in summer and least in spring. Lowest SAIDI values are observed in spring, and the total spread of values is greater in spring than in winter and autumn (omitting the outlier 2/04/08)

We can examine seasonal variation by year, and this is shown in Figure 10.

[5] Reminder about boxplots: Boxplots are a powerful shorthand way of visualising a distribution of values. The central line in the box represents the median (ie middle) value. The central box represents the middle 50% of values, so that the box ranges from the 25th to the 75th percentile. The 'whiskers' show the extent of most of the rest of the data, with extreme observations being represented by the outlying bars. (The whiskers extend as far as the largest (or smallest) observation lying within 1.5 × the length of the box.)

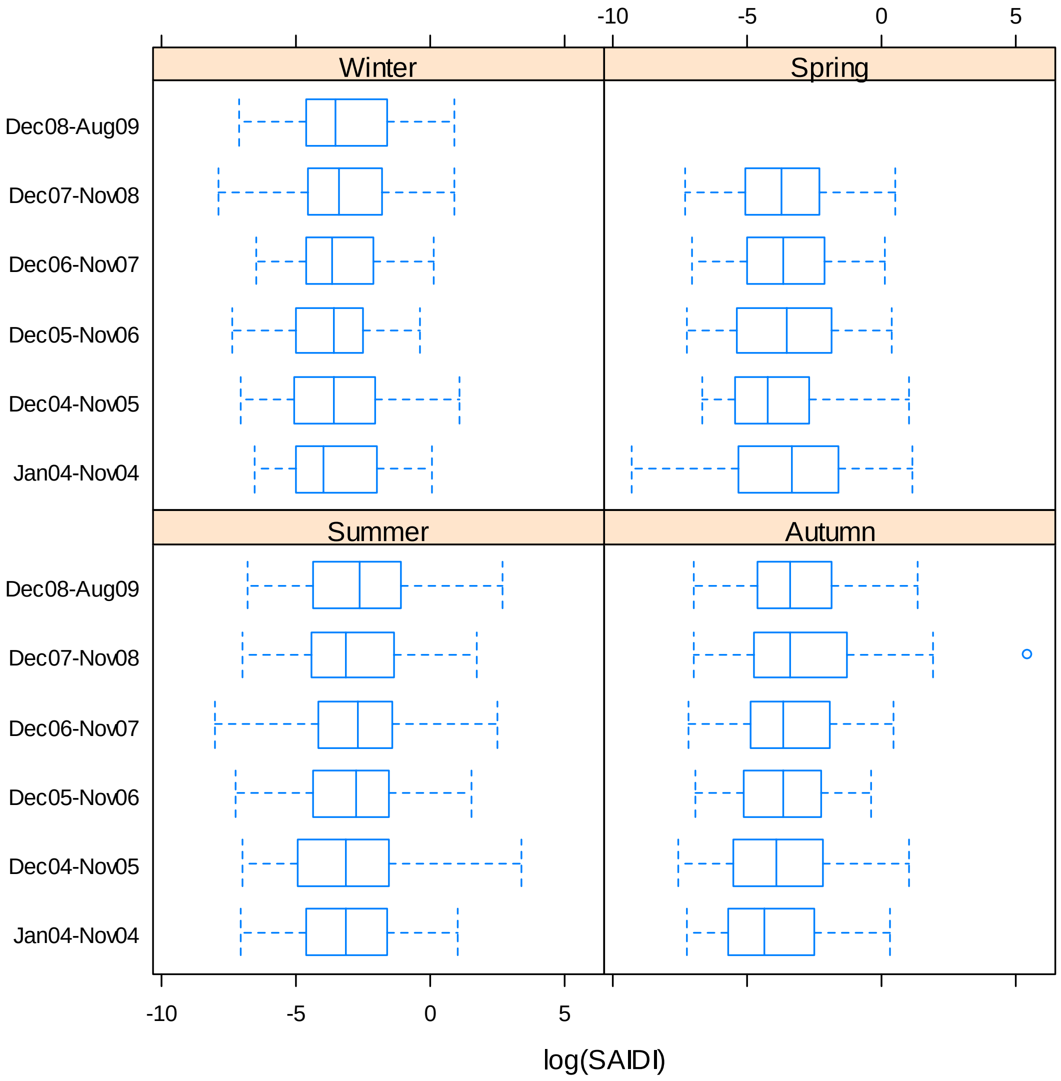

Figure 10: Seasonal variation by year

The medians of the distributions for summer, winter and spring are approximately stable, but the median for autumn has steadily increased over the period covered by the data. The values for the yearly medians are given below; the results are given as exp(median) ie equivalent to minutes per day, since it is an easier scale to assess.

Table 9: exp(median(log(SAIDI))) by season and year

Year	Summer	Autumn	Winter	Spring
Jan04-Nov04	0.045	0.012	0.019	0.035
Dec04-Nov05	0.043	0.020	0.028	0.014
Dec05-Nov06	0.064	0.025	0.028	0.029
Dec06-Nov07	0.068	0.025	0.026	0.025
Dec07-Nov08	0.044	0.033	0.034	0.024
Dec08-Aug09	0.072	0.033	0.030	-

On the log scale, there is a significant regression of the median on year for autumn (ie the median increases significantly with time), but not for the other seasons, although winter comes close to it.

We can also look at the normality of the distribution by season and year. Table 10 gives the results for the Anderson-Darling significance test.

Table 10: Significance tests by season and year

Year	Summer	Autumn	Winter	Spring	All
Jan04-Nov04	*P=0.04*	*P=0.04*	*P=0.003*	*P=0.01*	*P=7x10^{-6}*
Dec04-Nov05	*P=0.01*	P=0.17	P=0.24	*P=0.01*	*P=2x10^{-6}*
Dec05-Nov06	*P=0.04*	P=0.20	P=0.56	*P=0.02*	*P=0.0002*
Dec06-Nov07	P=0.56	P=0.12	*P=0.01*	*P=0.01*	*P=0.0002*
Dec07-Nov08	P=0.19	*P=0.01*	P=0.56	P=0.23	*P=0.003*
Dec08-Aug09	P=0.13	P=0.51	*P=0.02*	-	*P=0.002*

P values shown in italics are less than 0.05, a commonly accepted significance level. Note that even where P values are significantly smaller than 0.05, they are still much greater than for the whole year.

That is, individual seasons in individual years have SAIDI values which are close to log-normally distributed.

4.1 Comments

It appears that the mixture of data from seasons and years is resulting in a SAIDI distribution which is not log-normally distributed, even though data from the one season and year can be regarded as approximately log-normally distributed.

There is evidence that the SAIDI level is increasing during autumn, and possibly during winter, but the summer and spring distributions are more stable.

Appendix: R code

```
#
# read data from clipboard: Date, SAIDI, Year, FinYear, SeasYear, Season
#    and then set up factors and extra variables
#
data<-read.delim("clipboard")
data$Year<-factor(data$Year)
data$Date<-as.Date(data$Date,format="%d/%m/%Y")
data$Month<-months(data$Date)
data$Season<-
ifelse(data$Month=="December"|data$Month=="January"|data$Month=="February"|data$
Month=="March","Summer",
       ifelse(data$Month=="June"|data$Month=="July"|data$Month=="August","Winter
",     ifelse(data$Month=="April"|data$Month=="May","Autumn","Spring")))
data$Season<-
       factor(data$Season,ordered=TRUE,levels=c("Summer","Autumn","Winter","Spri
ng"))
attach(data)
#
# Set up conditions for calendar and financial years
#
YearCond<- Year!="2009"
FinYearCond<- FinYear!="03/04 (part)"&FinYear!="09/10 (part)"
#
# Figures 1&2 (and Figs 3&4 by substituting FinYearCond for YearCond)
#
require(lattice)
histogram(~log(SAIDI[YearCond]),xlab="log(SAIDI)")
require(car)
qq.plot(log(SAIDI[YearCond]),ylab="log(SAIDI)",line="robust",cex=0.5,
       col="dodgerblue")
#
#  Table 1 (and Table 3 by substituting FinYearCond for YearCond)
#
mean(log(SAIDI[YearCond])
median(log(SAIDI[YearCond])
skew<-function(x){  # set up function to calculate skewness & kurtosis & ci’s
       n<-length(x)
       require(e1071)
       g1<-skewness(x,type=1)
       sd.g1<-sqrt(6*(n-2)/((n+1)*(n+3)))
       cat("skewness:",g1," 95%ci: ",g1-1.96*sd.g1,g1+1.96*sd.g1,"\n\n")
       g2<-kurtosis(x,type=1)
       sd.g2<-sqrt(24*n*(n-2)*(n-3)/((n+1)*(n+1)*(n+3)*(n+5)))
       cat("kurtosis:",g2," 95%ci: ",g2-1.96*sd.g2,g2+1.96*sd.g2,"\n\n")
}
skew(log(SAIDI[YearCond]))

#
# Table 2 (and Table 4 by substituting FinYearCond for YearCond)
#
require(nortest)
ad.test(log(SAIDI[YearCond]))
lillie.test(log(SAIDI[YearCond]))
cvm.test(log(SAIDI[YearCond]))
shapiro.test(log(SAIDI[YearCond]))
#
# Box-Cox transformations (and similarly for FinYearCond)
#
box.cox.powers(SAIDI[YearCond])
lambda<-box.cox.powers(SAIDI[YearCond])$lambda
SAIDI.bcy<-box.cox(SAIDI[YearCond],lambda)
histogram(~SAIDI.bcy,xlab="Box-Cox(SAIDI)",main="Box-Cox(SAIDI), 2004-2008")
qq.plot(SAIDI.bcy,ylab="Box-Cox(SAIDI)",line="robust",cex=0.5,
       col="dodgerblue",xlab="Normal quantiles")
mean(SAIDI.bcy)
```

```
median(SAIDI.bcy)
skew(SAIDI.bcy)
#
# Yearly variation
#
qqmath(~log(SAIDI),groups=Year[,drop=TRUE],auto.key=list(corner=c(0.05,0.95),
       border=TRUE,pch=1),cex=0.5,data=data[YearCond,])
qqmath(~log(SAIDI),groups=FinYear[,drop=TRUE],auto.key=list(corner=c(0.05,0.95),
       border=TRUE,pch=1),cex=0.5,data=data[FinYearCond,])
bwplot(Season~log(SAIDI),horizontal=TRUE,pch="|",box.ratio=2)
bwplot(SeasYear~log(SAIDI)|Season,horizontal=TRUE,pch="|",box.ratio=2)
for(dd in levels(SeasYear))
       cat(dd,median(log(SAIDI[SeasYear==dd & Season=="Summer"])),"\n")
for(dd in levels(SeasYear))
       cat(dd,median(log(SAIDI[SeasYear==dd & Season=="Autumn"])),"\n")
for(dd in levels(SeasYear))
       cat(dd,median(log(SAIDI[SeasYear==dd & Season=="Winter"])),"\n")
for(dd in levels(SeasYear))
       cat(dd,median(log(SAIDI[SeasYear==dd & Season=="Spring"])),"\n")
```