New Action 0203: Jason Blackmore (JB) to provide more information on a formal validation check of the stratification approach for EUC Bands 01 and 02 for discussion at the April meeting.

During the February DESC meeting we discussed the issues with data quality from the 3rd party sources and impacts on the sample. Given the reduction in the size of the Xoserve managed sample and the increased availability of 3rd party data we expect greater variability between the characteristics of the current sample and new data sources. However, we have no way of quantifying or validating this until the full cycle of demand estimation analysis is completed.

Recommendation: To explore the use of Population Stability Index (PSI) as a validation measure for assessing the change in distribution between samples.

Apply PSI to compare the composition of the demand estimation sample against the population from all industry data. This may be informative to determine if the sample is representative or not and needs more focus.

Population Stability Index (PSI)

We use existing validation measures for accuracy (R2, MAPE) and the sample size formula to determine an optimum sample size for a given sampling error but lack a validation measure that can be used to determine the significant of any difference between the distribution of samples.

A validation measure Population Stability Index (PSI), can be used as a formal validation to determine if the distribution between two samples is significantly different, either changing over time or is different between samples.

The PSI measure has values that range from 0 to 1 and can be interpreted as follows:

PSI Values	Interpretation
0	No difference between the new and original samples
>0 – 0.25	Little difference
<0.25	Large difference – new sample is significantly different
1	Completely different

A PSI value >0.25 informs us that the resultant analysis could be more impacted by the addition of the new sample data as its distribution is significantly more different. Note, the PSI measure doesn’t inform us if the analysis will be accurate, just that the distribution of the sample is significantly different.

If there are already doubts about the quality of the data, then this measure could be used to determine if we want to include or exclude the new sample. Given an approach of using the most recent data the PSI can inform us if the analysis is likely to be significantly impacted by including the new sample.

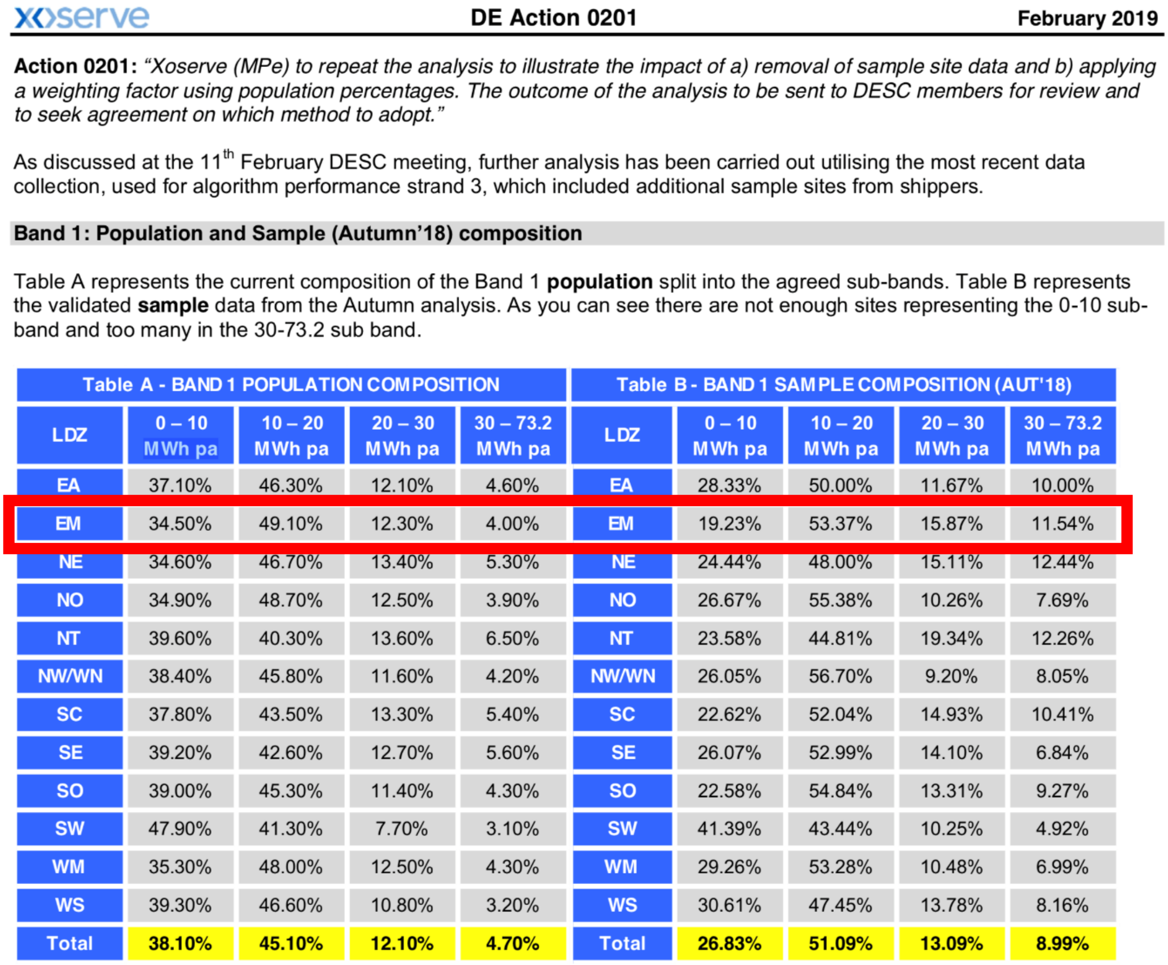
Example Use Case

During the February 2019 DESC meeting we discussed the impact of removing data to ensure that the sample more closely matched the distribution of the population – sample stratification. We apply sample stratification to ensure that the sample becomes more representative of the distribution of the population, thus becomes more accurate, but there was a trade-off as data was removed, thus increasing sample size error.

An example PSI calculation for LDZ EM is shown below using population and sample data from: https://gasgov-mst-files.s3.eu-west-1.amazonaws.com/s3fs-public/ggf/2019-02/1.4%20Action0201_Stratification.pdf

We can see it was noted that the Autumn 2018 sample (Table B) did not have enough sites representing the 0-10 MWh pa sub band and an overrepresentation in the 30-73.2 MWh pa sub band. The PSI measure can help determine if this difference is significant.

The PSI measure for EM is 0.18 – which is interpreted as little difference. Therefore, although a difference was noted in the distribution between the population and sample it could be not significant.



xoserve DE Action 0201 February 2019

Action 0201: *"Xoserve (MPe) to repeat the analysis to illustrate the impact of a) removal of sample site data and b) applying a weighting factor using population percentages. The outcome of the analysis to be sent to DESC members for review and to seek agreement on which method to adopt."*

As discussed at the 11th February DESC meeting, further analysis has been carried out utilising the most recent data collection, used for algorithm performance strand 3, which included additional sample sites from shippers.

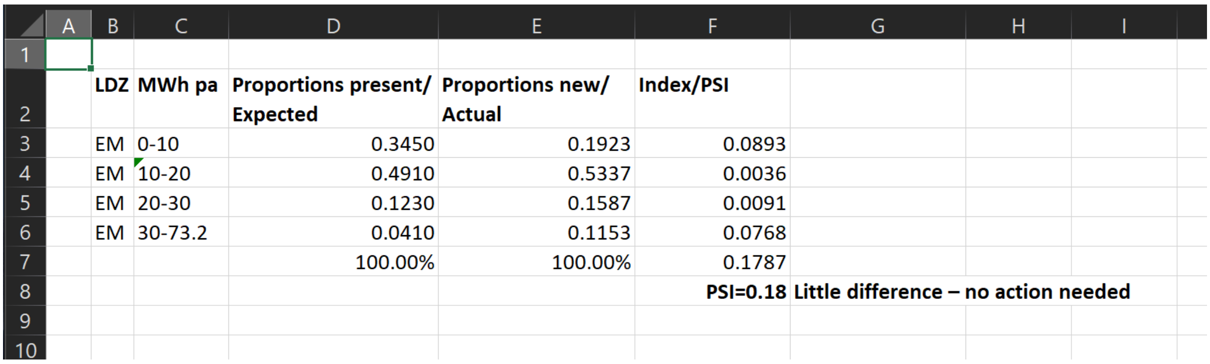
Band 1: Population and Sample (Autumn'18) composition

Table A represents the current composition of the Band 1 **population** split into the agreed sub-bands. Table B represents the validated **sample** data from the Autumn analysis. As you can see there are not enough sites representing the 0-10 sub-band and too many in the 30-73.2 sub band.

Table A - BAND 1 POPULATION COMPOSITION					Table B - BAND 1 SAMPLE COMPOSITION (AUT'18)				
LDZ	0 – 10 MWh pa	10 – 20 MWh pa	20 – 30 MWh pa	30 – 73.2 MWh pa	LDZ	0 – 10 MWh pa	10 – 20 MWh pa	20 – 30 MWh pa	30 – 73.2 MWh pa
EA	37.10%	46.30%	12.10%	4.60%	EA	28.33%	50.00%	11.67%	10.00%
EM	34.50%	49.10%	12.30%	4.00%	EM	19.23%	53.37%	15.87%	11.54%
NE	34.60%	46.70%	13.40%	5.30%	NE	24.44%	48.00%	15.11%	12.44%
NO	34.90%	48.70%	12.50%	3.90%	NO	26.67%	55.38%	10.26%	7.69%
NT	39.60%	40.30%	13.60%	6.50%	NT	23.58%	44.81%	19.34%	12.26%
NW/WN	38.40%	45.80%	11.60%	4.20%	NW/WN	26.05%	56.70%	9.20%	8.05%
SC	37.80%	43.50%	13.30%	5.40%	SC	22.62%	52.04%	14.93%	10.41%
SE	39.20%	42.60%	12.70%	5.60%	SE	26.07%	52.99%	14.10%	6.84%
SO	39.00%	45.30%	11.40%	4.30%	SO	22.58%	54.84%	13.31%	9.27%
SW	47.90%	41.30%	7.70%	3.10%	SW	41.39%	43.44%	10.25%	4.92%
WM	35.30%	48.00%	12.50%	4.30%	WM	29.26%	53.28%	10.48%	6.99%
WS	39.30%	46.60%	10.80%	3.20%	WS	30.61%	47.45%	13.78%	8.16%
Total	38.10%	45.10%	12.10%	4.70%	Total	26.83%	51.09%	13.09%	8.99%

Calculation

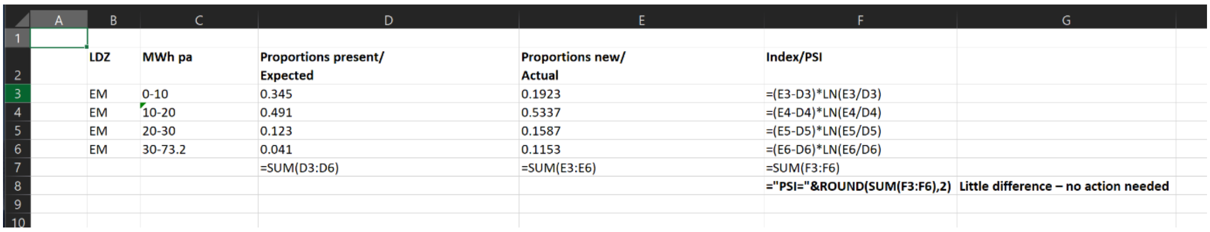
The calculation of PSI is shown below, for sub-band 0-10 MWh pa we can see the Index/PSI is a high proportion (0.0893), thus intuitively this makes sense as the calculation measures the distance between the distributions of the population and sample.



	A	B	C	D	E	F	G	H	I
1									
2		**LDZ**	**MWh pa**	**Proportions present/ Expected**	**Proportions new/ Actual**	**Index/PSI**			
3		EM	0-10	0.3450	0.1923	0.0893			
4		EM	10-20	0.4910	0.5337	0.0036			
5		EM	20-30	0.1230	0.1587	0.0091			
6		EM	30-73.2	0.0410	0.1153	0.0768			
7				100.00%	100.00%	0.1787			
8						**PSI=0.18**	**Little difference – no action needed**		
9									
10									

rounding has been applied to ensure proportions sum to 100%.

Formulas



	A	B	C	D	E	F	G
1							
2		**LDZ**	**MWh pa**	**Proportions present/ Expected**	**Proportions new/ Actual**	**Index/PSI**	
3		EM	0-10	0.345	0.1923	=(E3-D3)*LN(E3/D3)	
4		EM	10-20	0.491	0.5337	=(E4-D4)*LN(E4/D4)	
5		EM	20-30	0.123	0.1587	=(E5-D5)*LN(E5/D5)	
6		EM	30-73.2	0.041	0.1153	=(E6-D6)*LN(E6/D6)	
7				=SUM(D3:D6)	=SUM(E3:E6)	=SUM(F3:F6)	
8						**="PSI="&ROUND(SUM(F3:F6),2)**	**Little difference – no action needed**
9							
10							

Use Cases

There are several ways to apply PSI:

1. Calculate PSI over time for each year's sample data.
2. Calculate PSI between data sources. For example, between the Xoserve managed sample and the 3rd party sample data and interpret such as:

PSI >0.25	**Significant issues with data? (quantified/subjective)**	**Conclusion**
Yes – significant difference	Yes – data issues/concerns	Supports excluding the data.
Yes – significant difference	No data issues	Supports including the data being aware that the results can be affected by change in the distribution of the samples
No – little difference	Yes – data issues/concerns	Depends, for example supports including the data if data issues are minimal if the increase sample size will reduce sample error, if applicable.
No – little difference	No data issues	Supports including the data being aware that the results will not be significantly be affected by change in the distribution of the samples.

3. A further application of PSI would be to compare the composition of the demand estimation sample against the population from all industry data. A PSI measure greater than 0.25 would indicate that our sample doesn’t reflect the population. This may be informative to determine if the sample is representative or not and needs more focus.

Recommendation

To explore the use of PSI as a validation measure for assessing the change in distribution between samples.

Apply PSI to compare the composition of the demand estimation sample against the population from all industry data. This may be informative to determine if the sample is representative or not and needs more focus.

The PSI measure is widely used when comparing the distributions between samples and over time. The PSI values and interpretation could be subject to experience in the credit risk domain where is widely used, and so different values may suit other domains, which can be determined from experience and use.