**Slide 1 -** Dear ladies and gentlemen, my name is Matus Cuper and today I will present you our proposed method for anomaly detection of load consumption using clustering and statistical analysis. We were working on this solution with my colleagues from Slovak University of Technology Marek Loderer and Viera Rozinajova.

**Slide 2 -** During anomaly detection we encounter multiple problems such as data evolution, novelty detection, noise, definition of anomaly or lack of labeled datasets. There are several approaches of identification of anomalies for instance classification, clustering or statistical analysis. The anomalies can be divided into local and global anomalies. A local anomaly could be a spike during a normal period. It would not dramatically change overall consumption, in contrast with global anomaly that impact consumer consumption. The global anomalies can be easily identified by static thresholds. Aim of our work is to prevent illegal consumption, to identify malfunctioning devices and to optimize energy distribution.

**Slide 3 -** Since we were working with unlabeled dataset, we mainly focused on unsupervised methods that are clustering or statistical based. These solutions already dealing with some of these problems that I mentioned earlier such as novelty detection or data evolution.

**Slide 4 -** The new method I was working on with my colleagues consists of these steps.

**Slide 5 -** In the next slide we can see visualization of mentioned process. At the beginning the profiles of customers are clustered, then scored and the profiles with the highest score are being further analyzed by seasonal hybrid ESD analysis. At the end we visualize the combination of score and results of analysis. Now lets discus each step separately.

**Slide 6 -** So the first step is clustering of time series. In order to prepare dataset for clustering we had to normalize time series by z-score, so that we could have compared similarity of every single instance and not by absolute distance between data points. Also dataset is split into workdays and non-workdays due to different curve of electricity consumption. Other parameters like length of sliding window, overlapping, step length and distance metric may differ with different dataset, so they should be determined experimentally.

**Slide 7 -** Based on the results of clustering of each sliding window, the two types of scores are being calculated. Instance score is computed for each instance and it describes distance from instance to medoid of given cluster and it penalizes more distant ones. On the other hand cluster score is the same for instances in given cluster and it divides clusters into major and minor ones that are also penalized.

**Slide 8 -** In this slide we can see that clusters are divided into minor and major based on their size and the instance score is shown by gray arrows from medoids.

**Slide 9 -** Customers can be selected by multiple approaches, which is also dependent from dataset. It is also possible to use accounting data or any other data, that provide an added value. Our dataset contains only time series of load consumptions so we chose interval of interquartile rule and visualization for customer selection.

**Slide 10 -** For visualization we chose heatmap representation of consumer's data. As we can see days with higher value of score are colored by red. The disadvantage is that value of score does not change during given week as I mentioned earlier. Based on the visualization we can manually detect any patterns such as one time or repetitive events.

**Slide 11 -** The next step is to prepare data for seasonal hybrid extreme Studentized deviation analysis by smoothing. This step eliminates numerous local anomalies, that are not relevant. After that seasonal hybrid ESD analysis is applied to preprocessed data of subset of customers. Extreme Studentized deviation is based on Grubb's test. Its modification seasonal hybrid ESD uses median absolute deviation in process of time series decomposition in order to make method more robust and be capable of identifications up to fifty percent of anomalies. Output of this process is anomalousness flag, that is smoothed in order to group dense intervals and smooth sparse ones.

**Slide 12 -** And the last step is to combine smoothed flags from seasonal hybrid ESD analysis with consumer scores computed from clustering. This step transforms binary flags into real numbers with lower granularity.

**Slide 13 -** The results can be visualized by previously mentioned heatmap on daily, weekly or monthly basis.

**Slide 14 -** Another representation of score is to change line color of load consumption according to the value of score.

**Slide 15 -** For evaluation we used dataset from Irish CER Smart Metering Project. Since we did not have labeled dataset, we created synthetic one upon which the experiments were executed. Synthetic dataset consist of small group randomly selected consumers. Anomalies were created by replacing consumption of two random days with consumption of another customer. Setup of parameters for clustering was chosen based on the results of experiments using cluster validation indexes. This setup may differ for different datasets. The results of our proposed method were compared to the results of these methods.

**Slide 16 -** We can use this visualization as an example for comparison of our proposed method with original seasonal hybrid ESD analysis. As we can see in the first graph, there are two anomalies, both placed on Wednesday. However we can see that except of these two cases, there are several other ones also marked as anomaly. This is something that we successfully cleared from the results, in order to significantly increase that precision.

**Slide 17 -** And here we can see comparison of our method to the others. Method K-nearest neighbors is achieving the best value of precision, on the other hand seasonal hybrid ESD excels in recall. The F one score, which describes overall efficiency, achieves the best value by our method.

**Slide 18 -** To summarize our solution, better results are achieved by combining time series clustering and seasonal hybrid ESD method. Computation of score replaces flags by degree of anomalousness and it is performed only on preselected subset, which reduces overall computation time. The method can be easily extended for online processing of streaming data by incremental processing.