Rapidly Trainable Large-Scale Probabilistic Heat Pump Load Forecasting: A Kernel Density Estimation Approach

Leo Semmelmann
Karlsruher Institut für Technologie (KIT)
Karlsruhe, Germany
leo.semmelmann@kit.edu

Tobias Brudermueller
ETH Zurich
Zurich, Switzerland
tbrudermuell@ethz.ch

Abstract

As the electrification of heating accelerates, accurate and scalable probabilistic forecasting of heat pump loads becomes increasingly important for grid operators, utilities, and other practitioners. While state-of-the-art methods such as quantile gradient boosting offer high forecast accuracy, they are often computationally intensive and difficult to scale for large datasets. To address this challenge, we propose a Kernel Density Estimation (KDE)-based approach for probabilistic load forecasting in households with heat pumps. Using real-world data from 1,193 Swiss households, we benchmark KDE against quantile gradient boosting regarding forecast accuracy—measured via pinball loss—and computational efficiency. Our results show that KDE performs slightly better in 8 out of 9 quantiles while reducing training time by 3500×, though at the cost of higher inference time due to its sampling-based quantile estimation.

CCS Concepts

• **Computing methodologies → Modeling and simulation, Machine learning**.

Keywords

Heat pumps, load forecasting, kernel density estimation

ACM Reference Format:
Leo Semmelmann and Tobias Brudermueller. 2025. Rapidly Trainable Large-Scale Probabilistic Heat Pump Load Forecasting: A Kernel Density Estimation Approach. In *The 16th ACM International Conference on Future and Sustainable Energy Systems (E-ENERGY '25), June 17–20, 2025, Rotterdam, Netherlands.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3679240.3734640

1 Introduction

To reduce carbon emissions, policymakers worldwide are promoting a transition from fossil fuel-based to electricity-based heating systems [5]. Heat pumps represent a mature and efficient technology for achieving this electrification of heating.

The transition from gas boilers to heat pumps imposes a substantial additional load on electricity grids [17]. It is therefore essential to provide policymakers, grid operators, and utility companies with robust tools to accurately model future electricity demand from households equipped with heat pumps. While earlier research has predominantly focused on point forecasts of electricity load, there has been a growing recognition of the importance of probabilistic forecasting in recent years [10]. Probabilistic forecasting aims to estimate future electricity loads across a range of probability levels—such as specific quantiles—thereby capturing the inherent uncertainty in demand more comprehensively than traditional point forecasts.

Probabilistic forecasts have become essential tools in modern power systems, enabling more robust decision-making for grid congestion management, unit commitment, and extreme weather risk mitigation [7]. Common approaches to probabilistic load forecasting include quantile regression methods [10], gradient boosting techniques [13], and neural networks [18]. While these methods have demonstrated strong predictive performance, they often entail high computational costs, particularly when applied to datasets with large sample sizes [12, 19]. As an alternative method, Kernel Density Estimation-based approaches (KDE) have attracted increasing interest. These methods aim to estimate Gaussian kernels from training data and subsequently generate samples from the estimated distribution to construct quantile forecasts [1].

Although KDE has been applied in load forecasting contexts, existing studies have primarily focused on conventional household electricity consumption and have neither specifically examined the computational efficiency of such methods, nor investigated their performance with respect to heat pump-related data [1, 6]. Hence, in this study, we evaluate both the predictive performance and computational costs of a KDE-based approach for probabilistic load forecasting in households equipped with heat pumps. Our analysis draws on a novel, large-scale dataset comprising 1,193 Swiss households with heat pumps, monitored over a period of 1,576 days. The approach is benchmarked against a Gradient Boosting Quantile Regression.

The remainder of the study is structured as follows. In Section 2, the applied methodology, benchmarked approach and evaluation metric is presented. In Section 3, the applied data is described. Finally, 4 presents the results, discusses its implications and sheds light on promising research avenues, while Section 6 concludes the paper.

2 Methodology

The following methodology section is divided into three parts. First, the suggested KDE-based algorithm to forecast individual loads of households with heat pumps is described. Second, the quantile gradient boosting method is briefly introduced as the benchmarking method. Third, the evaluation metric is presented.

Kernel Density Estimation: The goal of kernel density estimation is to construct a nonparametric estimate of the density function f based on a set of observations, which, in our case, are hourly

E-ENERGY '25, Rotterdam, Netherlands

ACM ISBN 979-8-4007-1125-1/25/06
https://doi.org/10.1145/3679240.3734640

household energy consumption measurements ($Y_1, Y_2, \ldots, Y_N$, where N is the number of observations) [1]. The resulting estimator is given by:

$$\hat{f}(y) = \sum_{t=1}^{N} K_{h_y}(Y_n - y), \tag{1}$$

where $\hat{f}(y)$ denotes the estimated density at point y, $K_{h_y}(\cdot) = K(\cdot/h_y)/h_y$ represents the kernel function scaled by the bandwidth h_y, and h_y is the bandwidth parameter controlling the smoothness of the estimate. The kernel function K acts as a localized weighting function, assigning higher weights to observations closer to y and lower weights to those farther away, with the bandwidth determining the rate at which this influence decays. We employ a Gaussian kernel for the density estimation and determine the bandwidth based on preliminary experiments at 0.1. While alternative kernels, such as the Epanechnikov kernel [8], and established bandwidth selection methods, such as Silverman's rule of thumb [9], are available, initial experimental evaluations indicated that these alternatives resulted in inferior performance.

Each observation from the underlying dataset can be represented as a tuple (Y_n, T_n), where Y_n denotes the household energy consumption and T_n the corresponding temperature measurement at observation n. We discretize the temperature space into 1°C intervals and construct a separate kernel density estimate $\hat{f}_b(y)$ for each temperature bucket b. Each $\hat{f}_b(y)$ is estimated using only the subset of observations for which the temperature falls within the corresponding bucket. To ensure statistical robustness and reduce noise, we consider only those buckets that contain at least 1,000 observations. We construct temperature-dependent buckets due to the strong correlation between heat pump loads and ambient temperature [15], whereas the remaining household load is treated stochastically. While both heat pump and household loads can exhibit temporal patterns—such as peak consumption in the morning and evening—these effects could be addressed by introducing time-dependent buckets as well. However, for the sake of conciseness, we leave this extension to future work.

To infer a probabilistic quantile load forecast, we consider a set of forecasted temperatures $\{\tilde{T}_1, \tilde{T}_2, \ldots, \tilde{T}_m\}$. For each forecasted temperature $\tilde{T}_i$, we identify the corresponding temperature bucket b_i and draw S samples from the associated kernel density estimate. These samples form the basis for estimating quantiles of the household load distribution conditioned on the forecasted temperatures.

Quantile Gradient Boosting: The previously introduced KDE approach is benchmarked against a Gradient Boosting-based method, which is commonly used in the literature as a baseline due to its robustness, reproducibility, and performance [21]. Gradient Boosting Trees can be extended to probabilistic forecasting by training separate models to predict conditional quantiles of the target variable [4, 13]. Instead of minimizing the mean squared error, the model is trained with a quantile loss function that targets specific quantile levels. This approach allows us to generate prediction intervals and express uncertainty directly [1]. We note that a standard quantile regression approach was also considered during initial experiments. However, due to its prohibitively high runtime on the large underlying dataset, we limit our benchmarking to Quantile Gradient Boosting.

For both the kernel density estimation and quantile gradient boosting approaches, the input observations from each household are individually normalized to the [0,1] range using min-max scaling, where the maximum value (i.e., the household's peak load) serves as the scaling reference. After model training and inference, the outputs are rescaled accordingly using the same household-specific peak load. This normalization strategy enables the application of the forecasting models to previously unseen households, provided that information on heat pump size (and respective peak loads) is available. As a result, the approach is particularly well-suited for use by aggregators or grid operators aiming to estimate demand profiles for new or unmonitored households.

Metric: To evaluate the accuracy of the probabilistic forecasts, we use the pinball loss, also referred to as quantile loss. Unlike standard error metrics such as the Mean Squared Error, which evaluate point forecasts, the pinball loss assesses the quality of quantile predictions by accounting for the asymmetry in over- and under-predictions [16]. For a given quantile level q, the pinball loss at observation n is defined as:

$$L_{q,n}(y_n, \hat{y}_n^q) = \begin{cases} (1-q)(\hat{y}_n^q - y_n), & \text{if } \hat{y}_t^q \geq y_n \\ q(y_n - \hat{y}_n^q), & \text{if } \hat{y}_n^q < y_n, \end{cases} \tag{2}$$

where y_n is the true value and $\hat{y}_n^q$ is the predicted q-th quantile at observation n. The loss function penalizes overestimations and underestimations differently: when the forecasted quantile exceeds the actual value, the penalty is scaled by $(1-q)$, and when it falls below, it is by q. This makes the pinball loss a suitable metric for evaluating the sharpness and calibration of probabilistic forecasts.

3 Case study

We apply the previously described methodology to a high-quality dataset of 1,193 Swiss households with heat pump installations, sourced from the open-source HEAPO dataset [2]. Since this dataset consists of real-world smart meter data, it provides high-resolution insights into household energy consumption, and all results presented in this study are fully reproducible. The original dataset includes a subset of 214 households where heat pump systems were optimized by experienced energy consultants during on-site visits, providing time series data before and after the consultation. To ensure comparability across households while maintaining a large sample size, we exclude these treatment group households and focus solely on the control group without intervention. For our analysis, we use 15-minute smart meter data on total active energy consumption (in kWh) but aggregate it to hourly values. This temporal resolution is commonly applied in load forecasting tasks, such as day-ahead electricity procurement or scheduling of flexibility potentials [14].

For the evaluation, a five-fold cross-validation is applied [11], which provides a robust estimate of model performance by reducing

[1]We follow the implementation provided by `scikit-learn`'s `GradientBoostingRegressor`, which supports quantile regression natively: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingRegressor.html

the variance associated with a single train-test split. This approach enhances the generalizability and reliability of the results[2].

4 Result

In this section, we compare the quality and runtime of the proposed KDE-based method with the Quantile Gradient Boosting approach. All computations were performed on a device equipped with an Apple M1 chip.

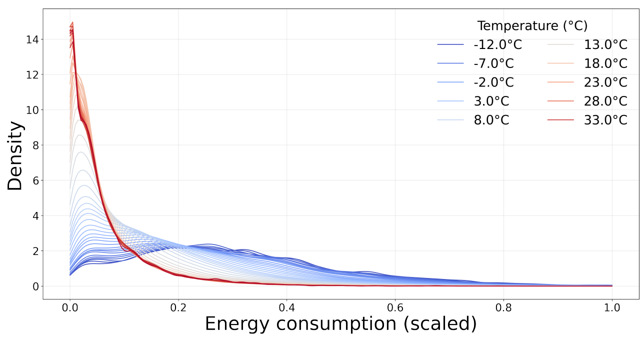

Figure 1: Kernel density estimates of scaled energy consumption across different temperature bins.

Figure 1 shows the temperature-dependent KDEs derived from the first k-fold training set[3]. As expected, the density shifts with temperature: at lower temperatures, the distributions are broader and peak at higher consumption levels, reflecting increased heating demand. In contrast, at moderate or higher temperatures, the density curves are narrower and concentrated around lower consumption values, indicating minimal or no heating-related electricity use across different households.

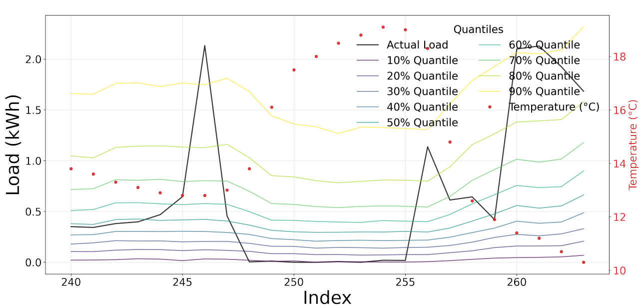

Figure 2: Exemplary KDE-based quantile forecasts.

For each temperature observation in the test sets, 1,000 sample values are drawn from the corresponding temperature-dependent KDE. These samples form the basis for the quantile estimates. An example of a KDE-based quantile forecast is shown in Figure 2.

[2]We note that in every cross-validation run, only up to 340,000 samples in the testing dataset were considered, due to computational reasons.

[3]While some density values exceed 1, this is based on the illustrative process. To validate the correctness of the KDEs, we numerically integrated each distribution, confirming that the area under each curve sums to 1.

Quantile Level	KDE	Gradient Boosting
10%	0.1141	**0.1140**
20%	**0.2250**	0.2250
30%	**0.3323**	0.3324
40%	**0.4355**	0.4357
50%	**0.5340**	0.5344
60%	**0.6262**	0.6267
70%	**0.7096**	0.7103
80%	**0.7787**	0.7795
90%	**0.8169**	0.8178

Table 1: Comparison of Pinball Loss between KDE and Gradient Boosting across quantiles. Lower values indicate better predictive accuracy for the respective quantile, with the best-performing method at each quantile shown in bold.

From a quality perspective, the KDE-based approach performs slightly better than the Gradient Boosting approach in 8 out of 9 quantiles, though the difference is not statistically significant ($p < 0.01$), as depicted in Table 1. The complete pinball loss values from all cross-validation runs are provided in the Appendix in Tables 3 and 2.

However, as illustrated in Figure 3, this advantage comes with significantly lower training times compared to the benchmark Quantile Gradient Boosting method. On average, training on 13,453,622 samples takes just 2 seconds using the KDE-based method, compared to 7,006 seconds with Gradient Boosting. In contrast, inference with the KDE approach requires 2,041 seconds, whereas Quantile Gradient Boosting completes inference in only 3 seconds. The higher inference time of the KDE-based method is primarily due to its dependence on the size of the test set, as quantiles for each entry are generated by sampling from the temperature-dependent KDEs. Later on, we discuss avenues to reduce inference time for the KDE-based method.

These findings have important implications for practitioners such as grid operators or utilities. While the KDE-based method offers significantly faster training—making it well-suited for rapid model updates on large datasets—it comes with higher inference costs due to its sampling-based quantile estimation. This trade-off makes the method especially appealing for use cases involving frequent retraining on large-scale data—such as thousands of households with fine-grained measurements—where only a limited number of future time steps need to be forecast, for instance day-ahead values.

5 Discussion

The previously presented results demonstrate that KDE-based forecasts are also suitable for large datasets of households with heat pump installations. While the general applicability of this method has already been established for standard household loads [1], it is important to confirm its effectiveness for households with heat pumps, as their load patterns change significantly and become heavily temperature-dependent due to the nature of heat pump operation [14].

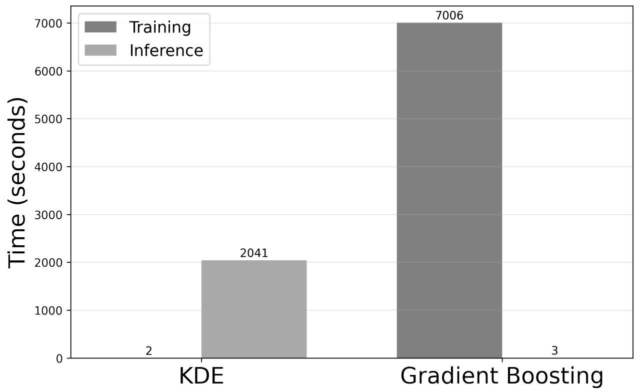

Figure 3: Runtimes for training and inference averaged over five-fold cross-validation runs.

Another key contribution of this work is the consideration of computation time in KDE-based probabilistic load forecasting. We show that the proposed approach yields reasonable forecasting performance for households with heat pumps while significantly reducing training time compared to the Gradient Boosting method. Nonetheless, several important aspects remain to be explored, presenting promising directions for future research.

First, we acknowledge that the KDE-based method was only benchmarked against Gradient Boosting. For a more comprehensive evaluation, it should also be assessed in terms of both performance and runtime against other approaches, such as neural networks [20].

Second, both methods were evaluated using only a single input variable—temperature measurements. Future studies should assess whether alternative approaches, such as the Quantile Gradient Boosting method, achieve superior performance when incorporating additional input features, such as humidity or time-of-day information. However, the KDE-based approach could also be extended to process additional inputs. One potential avenue is a conditional KDE approach [9], or alternatively, expanding the KDE estimation by introducing bins not only for temperature but also for factors like weekdays/weekends and different times of the day [1].

Third, this study did not fully exploit the granularity of the underlying dataset. Different temperature-dependent KDEs could have been fitted for various heat pump and building types. Additionally, while in our study measurements were resampled to hourly intervals, the dataset allows for finer time resolutions. These aspects present opportunities for further research and potential improvements to the proposed method.

Fourth, the rather basic implementation of the KDE in this study offers further room for refinement. The essential bandwidth parameter could be selected using more advanced parameter selection techniques, and alternative kernel functions could be explored. Additionally, boundary effects, which cause biased estimations at the lower and upper ends of the observed loads, could be mitigated through correction approaches, such as bandwidth adjustments near the boundaries [1, 3].

Fifth, we acknowledge the relatively high inference times of the KDE-based method, which may pose challenges for real-world applications. However, the current implementation has not incorporated any runtime optimizations. For example, in its current form, the model draws a large number of samples from the KDE for each individual test instance to construct quantiles. This process could be significantly accelerated by caching or reusing quantiles for similar temperature values, potentially leading to substantial reductions in inference time.

Overall, this study highlights the potential of KDE-based probabilistic forecasting for heat pump loads, demonstrating its fast training times. This serves as a foundation for future research and offers a promising method for both practitioners and researchers.

6 Conclusion

This study demonstrates that KDE offers a computationally efficient alternative to common approaches—such as Quantile Gradient Boosting—for probabilistic heat pump load forecasting, particularly due to its exceptionally fast training time. The KDE-based approach achieves comparable or slightly better accuracy in 8 out of 9 quantiles while reducing training time by 3500x, at the cost of higher inference times. These improvements make the approach particularly well-suited for large-scale forecasting tasks that demand frequent retraining but only require a small set of points to be forecast.

Despite the advantages of the KDE method, several areas remain for further research. The method was benchmarked only against Gradient Boosting and should be compared with other probabilistic forecasting techniques, such as neural networks. Additionally, only temperature was used as an input variable, and future work should explore whether alternative approaches benefit from incorporating additional factors like humidity or time-of-day effects. Moreover, while KDE bins were structured around temperature, refining the method by incorporating more granular clustering, such as building types or operational schedules, could improve its performance. Finally, KDE implementation can be enhanced by optimizing bandwidth selection, mitigating boundary effects to improve forecast accuracy and steps to speed up the inference process.

References

[1] Siddharth Arora and James W Taylor. 2016. Forecasting electricity smart meter data using conditional kernel density estimation. *Omega* 59 (2016), 47–59.

[2] Tobias Brudermueller, Elgar Fleisch, Marina González Vayá, and Thorsten Staake. 2025. HEAPO – An Open Dataset for Heat Pump Optimization with Smart Electricity Meter Data and On-Site Inspection Protocols. arXiv:2503.16993 [cs.CY] https://arxiv.org/abs/2503.16993

[3] Jing Dai and Stefan Sperlich. 2010. Simple and effective boundary correction for kernel densities and regression with an application to the world income and Engel curve estimation. *Computational Statistics & Data Analysis* 54, 11 (2010), 2487–2497.

[4] Jerome H Friedman. 2002. Stochastic gradient boosting. *Computational statistics & data analysis* 38, 4 (2002), 367–378.

[5] Ankita Singh Gaur, Desta Z Fitiwi, and John Curtis. 2021. Heat pumps and our low-carbon future: A comprehensive review. *Energy Research & Social Science* 71 (2021), 101764.

[6] Stephen Haben and Georgios Giasemidis. 2016. A hybrid model of kernel density estimation and quantile regression for GEFCom2014 probabilistic load forecasting. *International Journal of Forecasting* 32, 3 (2016), 1017–1022.

[7] Sue Ellen Haupt, Mayte Garcia Casado, Michael Davidson, Jan Dobschinski, Pengwei Du, Matthias Lange, Timothy Miller, Corinna Mohrlen, Amber Motley, Rui Pestana, et al. 2019. The use of probabilistic forecasts: Applying them in theory and practice. *IEEE Power and Energy Magazine* 17, 6 (2019), 46–57.

[8] Hui He, Junting Pan, Nanyan Lu, Bo Chen, and Runhai Jiao. 2020. Short-term load probabilistic forecasting based on quantile regression convolutional neural network and Epanechnikov kernel density estimation. *Energy Reports* 6 (2020), 1550–1556.

[9] Michael P Holmes, Alexander G Gray, and Charles Lee Isbell Jr. 2007. Fast nonparametric conditional density estimation. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*. 175–182.

[10] Tao Hong and Shu Fan. 2016. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting* 32, 3 (2016), 914–938.

[11] Yoonsuh Jung. 2018. Multiple predicting K-fold cross-validation for model selection. *Journal of nonparametric statistics* 30, 1 (2018), 197–215.

[12] Hee-Seok Oh, Thomas CM Lee, and Douglas W Nychka. 2011. Fast nonparametric quantile regression with arbitrary smoothing methods. *Journal of Computational and Graphical Statistics* 20, 2 (2011), 510–526.

[13] Priyesh Saini and SK Parida. 2024. A novel probabilistic gradient boosting model with multi-approach feature selection and iterative seasonal trend decomposition for short-term load forecasting. *Energy* 294 (2024), 130975.

[14] Leo Semmelmann, Matthias Hertel, Kevin J Kircher, Ralf Mikut, Veit Hagenmeyer, and Christof Weinhardt. 2024. The impact of heat pumps on day-ahead energy community load forecasting. *Applied Energy* 368 (2024), 123364.

[15] Leo Semmelmann, Patrick Jaquart, and Christof Weinhardt. 2023. Generating synthetic load profiles of residential heat pumps: a k-means clustering approach. *Energy Informatics* 6, Suppl 1 (2023), 37.

[16] Yi Wang, Dahua Gan, Mingyang Sun, Ning Zhang, Zongxiang Lu, and Chongqing Kang. 2019. Probabilistic individual load forecasting using pinball loss guided LSTM. *Applied Energy* 235 (2019), 10–20.

[17] SD Watson, J Crawley, KJ Lomas, and RA Buswell. 2023. Predicting future GB heat pump electricity demand. *Energy and Buildings* 286 (2023), 112917.

[18] Lei Xu, Shengwei Wang, and Rui Tang. 2019. Probabilistic load forecasting for buildings considering weather forecasting uncertainty and uncertain peak load. *Applied energy* 237 (2019), 180–195.

[19] Jiyan Yang, Xiangrui Meng, and Michael Mahoney. 2013. Quantile regression for large-scale applications. In *International Conference on Machine Learning*. PMLR, 881–887.

[20] Wenjie Zhang, Hao Quan, and Dipti Srinivasan. 2018. An improved quantile regression neural network for probabilistic load forecasting. *IEEE Transactions on Smart Grid* 10, 4 (2018), 4425–4434.

[21] Wenjie Zhang, Hao Quan, and Dipti Srinivasan. 2018. Parallel and reliable probabilistic load forecasting via quantile regression forest and quantile determination. *Energy* 160 (2018), 810–819.

Appendix

Table 2: Pinball Loss per quantile with KDE-based approach, per every cross-validation fold.

Fold	Pinball_Loss_10	Pinball_Loss_20	Pinball_Loss_30	Pinball_Loss_40	Pinball_Loss_50	Pinball_Loss_60	Pinball_Loss_70	Pinball_Loss_80	Pinball_Loss_90
1	0.1084	0.2137	0.3155	0.4133	0.5064	0.5932	0.6711	0.7344	0.7665
2	0.1218	0.2408	0.3562	0.4676	0.5744	0.6751	0.7675	0.8459	0.8940
3	0.1071	0.2114	0.3121	0.4089	0.5011	0.5873	0.6651	0.7291	0.7643
4	0.1198	0.2362	0.3487	0.4570	0.5603	0.6570	0.7446	0.8169	0.8563
5	0.1131	0.2228	0.3289	0.4307	0.5276	0.6181	0.6998	0.7673	0.8035
Average	0.1141	0.2250	0.3323	0.4355	0.5340	0.6262	0.7096	0.7787	0.8169

Table 3: Pinball Loss per quantile with Quantile Gradient Boosting approach, per every cross-validation fold.

Fold	Pinball_Loss_10	Pinball_Loss_20	Pinball_Loss_30	Pinball_Loss_40	Pinball_Loss_50	Pinball_Loss_60	Pinball_Loss_70	Pinball_Loss_80	Pinball_Loss_90
1	0.1084	0.2138	0.3156	0.4135	0.5068	0.5938	0.6718	0.7352	0.7675
2	0.1218	0.2408	0.3563	0.4678	0.5747	0.6756	0.7681	0.8467	0.8949
3	0.1071	0.2114	0.3122	0.4091	0.5015	0.5878	0.6657	0.7299	0.7651
4	0.1198	0.2362	0.3488	0.4572	0.5607	0.6576	0.7453	0.8178	0.8572
5	0.1131	0.2229	0.3290	0.4310	0.5281	0.6187	0.7005	0.7681	0.8044
Average	0.1140	0.2250	0.3324	0.4357	0.5344	0.6267	0.7103	0.7795	0.8178