Filling the Gap in Time: Intelligent Imputation of Historical Parish Records

Adam Kiersztyn
Lublin University Of Technology; Department of Computational Intelligence
Lublin, Poland *a.kiersztyn@pollub.pl*

Piotr Rachwał
The John Paul II Catholic University of Lublin; Department of Source Studies Archvistics and History Didactics
Lublin, Poland *piotr.rachwal@kul.pl*

Krystyna Kiersztyn
Lublin University Of Technology; Department of Computational Intelligence
Lublin, Poland *k.kiersztyn@pollub.pl*

Abstract

The paper presents a method for imputing missing monthly values in historical parish records, based on data from two Prussian towns (Stargard and Słupsk, 1886–1913). The approach combines 12 simple estimation methods with predictive models, including Random Forest and Gradient Boosted Trees. The evaluation was performed using synthetic gaps introduced into complete datasets, with each experiment repeated independently 10 times. The results show that aggregation models significantly reduce the relative error of imputation, with ensemble models achieving average errors below 1%. The solution is general and can be adapted to similar historical sources. Potential applications include demographic series reconstruction and the detection of anomalies in archival data. .

Keywords: missing data imputation, historical demography, parish records, statistical modeling, AI techniques.

1. Introduction

Historical demography relies heavily on vital statistics to reconstruct the size, structure, and dynamics of past populations. These records—baptisms (births), marriages, and burials (deaths) — were traditionally maintained by religious institutions and later by civil authorities. Among them, parish registers stand out as indispensable sources, particularly in the context of pre-statistical societies.

However, these historical sources are often incomplete. The causes range from wars, epidemics, and negligence to physical deterioration of the record books. This problem is especially prominent in smaller towns and parishes, where the continuity of registration was more fragile. Missing data severely hinders the ability to conduct long-term, quantitative analyses of demographic patterns, such as fertility, mortality, or population growth.

This article focuses on addressing such gaps using data exclusively from two urban centers in historical Prussia: Stargard and Słupsk, covering the years 1886–1913. The monthly figures for births and deaths come from the Imperial Health Office's statistical publications, which recorded data for cities with more than 15,000 inhabitants.

We propose an adaptable, multi-method approach to imputing missing monthly data. This method combines traditional statistical tools (e.g., local means, seasonal medians) with artificial intelligence techniques such as Random Forest and Gradient Boosted Trees. The goal is to improve the reliability of incomplete data series and make them usable for demographic research.

Through extensive numerical experiments, we show that AI-based aggregation of multiple imputation techniques significantly reduces estimation error—often below 1%. This confirms that the method is not only effective but also scalable, interpretable, and applicable to other similar historical datasets.

Filling in data gaps alongside anomaly detection has become one of the key challenges in the field of data analysis and machine learning [4, 5, 6, 7, 8, 9, 10]. Many collected datasets contain missing values, which can result from various factors such as measurement errors, technical issues, or simply lack of information. These gaps can negatively impact the quality of analysis and modeling, leading to incomplete and inappropriate results. Data gap filling plays a crucial role in the process of data analysis and interpretation [3]. There are numerous techniques and methodologies that can be applied to address this problem, depending on the type of data, available information, and analysis objectives [13]. Earlier studies proposed various imputation techniques, including statistical substitution [10], multiple imputation [12], and matrix completion methods. More recent approaches include deep learning [13] and fuzzy rule-based systems [11]. This work differs by combining multiple simple methods through interpretable models.

2. Data and Structure

The dataset used in this study consists of monthly aggregated records of births and deaths in two historical towns of Prussia: Stargard and Słupsk. These towns were included in the statistical bulletins published by the Imperial Health Office in Berlin between 1886 and 1913. The records were originally presented in the form of monthly tables for each year, covering basic vital statistics collected in urban municipalities with more than 15,000 inhabitants.

The structure of the dataset can be described as a time series matrix, where each row corresponds to a specific month (from January to December), and each column represents a year within the range of the available period. For both locations, the matrix includes two separate series: one for births and one for deaths. The entries contain non-negative integers indicating the number of events observed in a given month.

Missing values were identified as empty cells in the dataset. Any cell that contained a valid number (including zero) was considered complete and excluded from imputation. Only true absences of data—where no value was recorded—were treated as gaps requiring estimation.

For the purpose of evaluation, artificial gaps were introduced into the otherwise complete dataset. Specifically, for each town and each type of event (births and deaths), five missing values were randomly inserted for every calendar month. This resulted in 60 artificial gaps per event type per town (12 months × 5 gaps), and a total of 240 gaps per town.

The gaps were inserted randomly across the entire time span (1886–1913), with equal probability for all years. The artificial removal of values allowed the authors to calculate relative errors by comparing the imputed values against the original data. This made it possible to assess the accuracy and robustness of each imputation method under controlled conditions. The procedure of random gap insertion and evaluation was repeated independently 10 times. All reported results represent average values across these runs.

As mentioned previously, the present method serves to fill in the data gaps in the monthly data series. Its main purpose is not to detect possible underestimations of the data entered, but only to fill the identified data gaps, which for the purposes of numerical experiments will be equated with the lack of an entry in the cell corresponding to a given month. If a certain value (even zero) is entered in a cell, the cell is not analysed. On the other hand, if, on the basis of the developed model, the results for the tested cell are zero, then it should be assumed that the data gap is caused by the actual lack of events in that month.

In order to present the operation and functionality of the method, monthly summaries of the data series of vital statistics from two Prussian towns were used. The series of vital statistics was used for two cities from the area of historical Prussia, i.e. Stargard Szczeciński (Stargard in Pommern) and Słupsk (Stolp). These were relatively large urban centers. In 1880 Słupsk was inhabited by 21557 people, in 1900 the number of inhabitants was 27,283, and in 1910 – 33762 . In Stargard Szczeciński these numbers were respectively: 21795, 26858, 27551. The monthly statements of the birth and death numbers for 1886 – 1913 were taken from the published Health Monitors issued by the Imperial Health Office in Berlin in the *Veröffentlichungen des Kaiserlichen Gesundheitsamts* series [2]. The source, in addition to the normative content, contains monthly data on the basic elements of the natural movement of the population and the causes of (sudden) deaths caused mainly by acute infectious diseases in urban and rural communes with 15 000 inhabitants and more. More information on the demographic analysis of selected urban centers in Pommern can be found in [1].

3. Imputation Methods

In total, twelve different methods were used to estimate missing values in the time series. These include simple statistical techniques, local smoothing procedures, and a cross-location method using correlated data from a second parish. The first method (M1) consists in calculating the average for a given month across the entire available period. In contrast, M2 replaces the missing value with the average for the entire year in which the gap occurs. M3 uses the year-to-year relative increase in the annual average, while M4 applies the same approach based on the median.

Methods M5 and M6 are based on the values of the months surrounding the missing entry. In M5, the missing value is estimated by the average of the month directly preceding and directly following the gap. M6 extends this approach to include the two months before and the two months after the missing point.

M7 and M8 use data from the same calendar month in neighboring years. M7 takes the average from two adjacent years, while M8 considers four years (two before and two after the year in which the gap appears). These approaches aim to preserve the seasonality of the data, which is often important in demographic series.

M9 and M10 use weighted averages based on a 3×3 mask applied over a grid of months and years. The weight matrices are defined in formulas (1) and (2), respectively. These methods rely on the assumption that the values of neighboring months and years can be used to infer the missing one, with greater weights assigned to closer entries.

M11 uses the median for the year in which the gap occurs. This method is less sensitive to outliers and may be more appropriate in case of irregular data.

Finally, M12 is based on the assumption of similarity between two parishes (in this case, Stargard and Słupsk). The monthly relative increase observed in one town is applied to the previous value in the other town to estimate the missing entry. This method is only applicable when parallel data from both towns are available and strongly correlated.

All of the above methods yield individual estimates, which are then used as input variables in the construction of predictive models aggregating the results.

4. Model Aggregation

The estimates obtained using individual imputation methods (M1–M12) were treated as explanatory variables in the construction of predictive models. The aim of this stage was to verify whether appropriate aggregation of simple estimation techniques improves the accuracy of imputation. Five models were considered for this purpose: linear regression without a constant term, linear regression with a constant, polynomial regression, Random Forest, and Gradient Boosted Trees.

Linear regression models were estimated using the ordinary least squares method. The first variant (Linear 1) assumed no intercept, while the second (Linear 2) included a constant term. The choice to include both versions in the study was dictated by the need to examine the effect of model specification on predictive accuracy. Polynomial regression was limited to the second degree, and only main effects (squared terms) were included, without interaction terms, to avoid overfitting. All linear and polynomial models were built using the STATISTICA software environment.

Models based on artificial intelligence techniques—Random Forest and Gradient Boosted Trees—were implemented in the KNIME Analytics Platform. In the Random Forest model, the number of trees was set to 100, and no limit was imposed on tree depth. Bootstrap sampling was enabled. For the Gradient Boosted Trees model, 100 iterations were used, with a learning rate of 0.1, and the maximum tree depth was fixed at 5. The models were trained using squared error loss. All other hyperparameters were left at their default settings.

The learning procedure was based on a random 80/20 split of the data. The models were trained on 80% of the cases (i.e. artificially generated missing values) and tested on the remaining 20%. This procedure was repeated separately for births and deaths and for each of the two towns. For each data point with a missing value, the model received 12 variables (results of M1–M12) as input and returned a single estimated value as output. This allowed direct comparison between model predictions and the true (known) value removed earlier in the process.

The main advantage of the proposed approach lies in the possibility of combining complementary strengths of individual estimation methods. While simple statistical techniques capture long-term or seasonal patterns, models such as Random Forest or Gradient Boosted Trees are able to detect non-linear dependencies and interactions between estimates. In addition, ensemble models are robust to outliers and do not require strong assumptions regarding the distribution of the data.

The predictive performance of all five models was assessed using the relative error of imputation. The results are presented and discussed in the following section.

5. Results and Comparison of Methods

Two statistical tools were used to determine the predictive models. Linear models were estimated using the popular statistical package STATISTICA, while the remaining models were developed using the KNIME program.

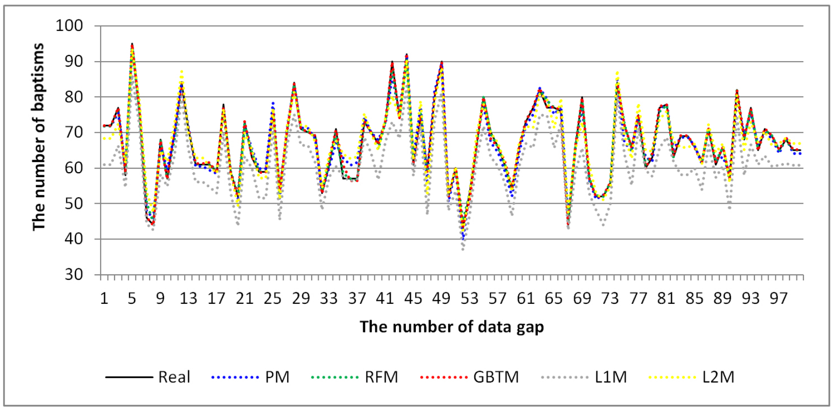

Fig. 1. Comparison of the effectiveness of models for the baptisms series in the parish of Stargard

By comparing the actual values of the number of births in the two Prussian urban centers with the results of the estimates obtained with the use of 5 models (see Figure 1 and Figure 2), we can see that the worst results are obtained when using a linear model without a constant value of α_0. The values of the basic statistics describing the relative error for the individual methods of filling the missing data, as well as for all five models describing the number of births for both

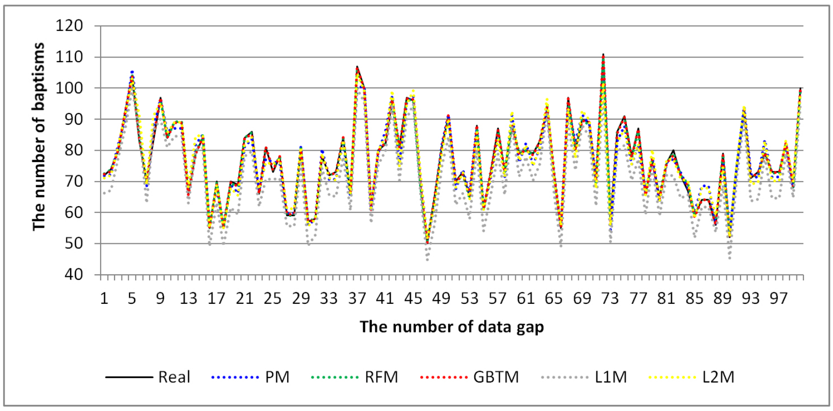

Fig. 2. Comparison of the effectiveness of models for the baptisms series in the Słupsk

analysed urban centers are presented in Table **??** and Table **??**.

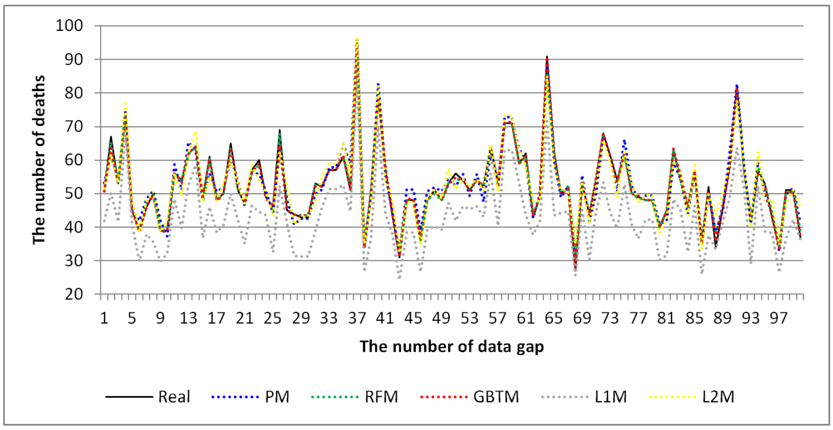

Fig. 3. Comparison of the effectiveness of the models for the series of deaths in the Słupsk

By comparing the graphs of the real data and estimates obtained with the use of individual predictive models, it may be noted that in the case of modelling the number of deaths (see Figure 3 and Figure 4), the best fit is provided by the models using AI techniques.

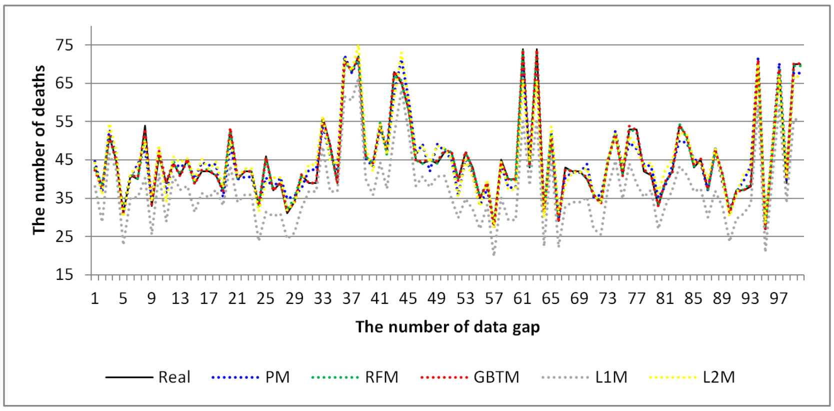

Fig. 4. Comparison of the effectiveness of models for deaths in Stargard

When analysing the values of the basic statistics for individual methods of filling in data gaps, as well as individual predictive models aggregating the results of individual methods, it may be seen that the average relative error as well as other positional statistics for this error prove the superiority of models using AI techniques. In the case of filling in artificially generated data gaps in the number of deaths in the Stargard (see Table 1), we can see that the average error for the model using Random Forest and Gradient Boosted Trees is less than 1%, which proves its high degree of efficiency. Moreover, it turns out that in more than half of the cases the relative error is of the order of 0.5%.

Table 1. Relative error statistics for deaths in the Stargard.

	Minimum	Q_1	**Average**	Median	Q_3	**Maximum**
Method 1	0.001	0.057	0.154	0.118	0.215	0.837
Method 2	0.003	0.084	0.188	0.157	0.268	0.757
Method 3	0.006	0.029	0.070	0.063	0.106	0.192
Method 4	0.011	0.064	0.099	0.103	0.143	0.188
Method 5	0	0.080	0.209	0.177	0.304	0.833
Method 6	0	0.080	0.205	0.183	0.293	0.765
Method 7	0	0.035	0.109	0.077	0.151	0.502
Method 8	0	0.054	0.132	0.104	0.174	0.846
Method 9	0	0.044	0.137	0.108	0.203	0.770
Method 10	0.001	0.042	0.136	0.105	0.203	0.747
Method 11	0	0.067	0.178	0.143	0.263	0.826
Method 12	0	0.088	0.222	0.176	0.279	1.652
Polynomial	0	0.017	0.044	0.036	0.064	0.170
Random forest	0	0.002	0.010	0.006	0.014	0.087
Gradient boosted trees	0	0.003	0.008	0.005	0.010	0.105
Linear 1	0.005	0.135	0.181	0.179	0.233	0.349
Linear 2	0	0.022	0.054	0.048	0.077	0.154

By examining the values of the basic statistics for the relative error obtained when estimating the number of deaths in the Słupsk (see. Table 2), we arrive at the conclusion that the appropriate aggregation of the results of individual methods of filling deficiencies is a key element of the proposed innovative solution. The use of such a simple model as the linear model with a constant α_0 facilitates the achievement of results in which the average error does not reach 6%.

The results obtained for birth records in both urban centres confirm the effectiveness of the proposed approach. In both Stargard and Słupsk, the lowest values of relative error were observed for models using artificial intelligence techniques. In particular, the Random Forest and Gradient Boosted Trees models returned the smallest average and median errors, with values frequently below 1%. Polynomial regression and the linear model with intercept also provided satisfactory results.

Among the individual methods, the best results were achieved using M3 and M4, based on the annual increase in the average or median, and M7 and M8, using the same month in two or four neighbouring years. These methods take into account seasonal variation and long-term trends. In contrast, methods relying on overall monthly or yearly averages (M1, M2) returned higher estimation errors.

The use of relative increases from the second parish (method M12) did not yield satisfactory results. The low level of agreement between the values recorded in both parishes appears to be the cause. In cases where a stronger correlation is present, this method may be more effective.

6. Discussion

Despite the high effectiveness of the proposed approach, several limitations should be noted. Method M12, based on relative monthly increases from the second parish, did not provide sat-

Table 2. Relative error statistics for deaths in the Słupsk.

	Minimum	Q_1	**Average**	Median	Q_3	**Maximum**
Method 1	0.001	0.066	0.173	0.127	0.227	0.792
Method 2	0.001	0.066	0.192	0.142	0.269	0.840
Method 3	0.003	0.062	0.095	0.089	0.142	0.260
Method 4	0.028	0.053	0.093	0.071	0.107	0.338
Method 5	0	0.082	0.227	0.188	0.315	1.038
Method 6	0	0.070	0.208	0.167	0.295	0.818
Method 7	0	0.046	0.116	0.094	0.147	1.025
Method 8	0	0.066	0.141	0.110	0.196	0.908
Method 9	0	0.050	0.148	0.131	0.211	0.817
Method 10	0	0.048	0.147	0.125	0.212	0.799
Method 11	0	0.073	0.186	0.156	0.266	0.759
Method 12	0	0.091	0.238	0.216	0.326	1.375
Polynomial	0	0.018	0.044	0.037	0.063	0.212
Random forest	0	0.002	0.012	0.006	0.014	0.228
Gradient boosted trees	0	0.002	0.008	0.005	0.009	0.099
Linear 1	0.003	0.139	0.197	0.203	0.253	0.364
Linear 2	0	0.022	0.056	0.047	0.075	0.390

isfactory results. The low correlation between the event patterns in Stargard and Słupsk limits the usefulness of this approach. In cases where a stronger relationship between parishes is confirmed, better results may be expected.

The evaluation was based on artificially generated gaps in otherwise complete series. Although this allows for direct error calculation, it does not reflect the actual distribution of missing data in historical sources. In practice, missing entries are often clustered or result from structural factors, which may affect the performance of individual methods.

It should also be emphasized that the results depend on the location of the missing data. Gaps occurring during periods of strong seasonal variation or sudden changes may lead to higher errors. This is confirmed by the variability of results across different random samples.

Finally, it must be stressed that imputed values are estimates and cannot be treated as original data. Although the relative error is low, there remains a degree of uncertainty. The use of reconstructed values in further historical analyses requires caution. Particular care is advised when interpreting short-term fluctuations or drawing conclusions based on rare or extreme events.

7. Conclusions and Future Work

The results obtained in the study confirm the high effectiveness of the proposed approach. The best outcomes were achieved using predictive models based on artificial intelligence techniques. Random Forest and Gradient Boosted Trees consistently returned the lowest values of relative error, often below 1%. Models based on linear and polynomial regression also provided satisfactory results, although with higher variability.

The use of multiple independent estimation methods and their aggregation in predictive models proved to be an effective strategy. The proposed solution is flexible and can be adapted to other historical datasets with similar characteristics.

Further work will focus on extending the method to a broader dataset covering 61 parish centres from Eastern Poland, with monthly data for the 18th and 19th centuries. The level of completeness and continuity in these sources varies, which will allow for testing the robustness of the method in different conditions.

Additional directions of research include the application of aggregation techniques based on fuzzy sets and experiments using fuzzy integrals, including the Choquet integral. Work is also underway on anomaly detection in historical data series, which may allow for the identification of underestimated or overestimated values in the source material. The method can be used in historical-demographic research requiring complete time series, such as fertility reconstruction, mortality seasonality, or crisis event detection.

Funding

The work was co-financed by the Lublin University of Technology Scientific Fund: FD-20/IT-3/002.

References

[1] Chojecki, D.: Od społeczeństwa tradycyjnego do nowoczesnego. Demografia i zdrowotność głównych ośrodków miejskich Pomorza Zachodniego w dobie przyspieszonej industrializacji i urbanizacji w Niemczech (1871-1913). Szczecin: Wydawnictwo Naukowe Uniwersytetu Szczecińskiego, 2014.

[2] Gesundheitsamt, D. R.: Veröffentlichungen des Kaiserlichen Gesundheitsamts. Berlin: Verlag von Julius Springer, 1918.

[3] Jerez, J. M., Molina, I., García-Laencina, P. J., Alba, E., Ribelles, N., Martín, M., and Franco, L.: Missing data imputation using statistical and machine learning methods in a real breast cancer problem. In: *Artificial intelligence in medicine* 50.2 (2010), pp. 105–115.

[4] Karczmarek, P., Kiersztyn, A., Pedrycz, W., and Al, E.: K-means-based isolation forest. In: *Knowledge-based systems* 195 (2020), p. 105659.

[5] Karczmarek, P., Kiersztyn, A., Pedrycz, W., Badurowicz, M., Czerwiński, D., and Montusiewicz, J.: K-medoids clustering and fuzzy sets for isolation forest. In: *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE. 2021, pp. 1–8.

[6] Karczmarek, P., Kiersztyn, A., Pedrycz, W., and Czerwiński, D.: Fuzzy c-means-based isolation forest. In: *Applied Soft Computing* 106 (2021), p. 107354.

[7] Kiersztyn, A., Czerwiński, D., Czermański, E., Oniszczuk-Jastrząbek, A., Smoliński, K., Miazek, P., Laskowicz, T., Jankiewicz, J., Rzepka, A., and Miśkiewicz, R.: Data integrity analysis on the example of AIS database. In: *Scientific Papers of Silesian University of Technology. Organization & Management/Zeszyty Naukowe Politechniki Slaskiej. Seria Organizacji i Zarzadzanie* 208 (2024).

[8] Kiersztyn, A., Karczmarek, P., Kiersztyn, K., and Pedrycz, W.: Detection and classification of anomalies in large datasets on the basis of information granules. In: *IEEE Transactions on Fuzzy Systems* 30.8 (2021), pp. 2850–2860.

[9] Kiersztyn, A., Karczmarek, P., Łopucki, R., Pedrycz, W., Al, E., Kitowski, I., and Zbyryt, A.: Data imputation in related time series using fuzzy set-based techniques. In: *2020 IEEE international conference on fuzzy systems (FUZZ-IEEE)*. IEEE. 2020, pp. 1–8.

[10] Little, R. J. and Rubin, D. B.: Statistical analysis with missing data. John Wiley & Sons, 2019.

[11] Pedrycz, W.: Fuzzy sets in pattern recognition: methodology and methods. In: *Pattern recognition* 23.1-2 (1990), pp. 121–146.

[12] Rubin, D. B.: *Multiple imputation for survey nonresponse*. 1987.

[13] Yoon, J., Jordon, J., and Schaar, M.: Gain: Missing data imputation using generative adversarial nets. In: *International conference on machine learning*. PMLR. 2018, pp. 5689–5698.