# Generic Data Imputation and Feature Extraction for Signals from Multifunctional Printers

Jakub Valcik*
Konica Minolta Laboratory Europe
Brno, Czech Republic
jakub.valcik@konicaminolta.cz

Wojciech Indyk
Konica Minolta Laboratory Europe
Brno, Czech Republic
wojciech.indyk@konicaminolta.cz

## ABSTRACT

Printer devices evolved into complex machines from both a hardware and software perspective. Nowadays, they are multifunctional and generate a variety of signals. The signals are mainly utilized for post-fact diagnosis of a printer. There is a great opportunity to employ the signals as inputs to decision support systems. Commonly, a decision support system requires a specific format and characteristic for the input data. In this paper, we analyze the characteristics of data generated by Multifunctional Printers, and compare and propose an optimal approach to handle missing signal values. To the best of our knowledge, this is the very first study that prepares a structured dataset for multifunctional printer signals. The proposed approach has been examined on a real-world dataset of signals of printers from Konica Minolta Inc.

## 1 INTRODUCTION

The evolution of business strategies from product-oriented to a service-oriented approach can be seen especially in IT companies [7] - IBM and Microsoft are mature examples of making this transition [2]. The IT sector is transparent and elastic to adopt new business models such as SaaS or transaction-based models in terms of revenue and remote web-based (cloud) or bundled as part of a hardware product in terms of delivery [7, 18]. Printing, a traditionally product-oriented market where printers and copying machines were sold to end-customers, also introduced the trend of delivering services [3]. Managed Print as a Service (MPS) frees the customer from taking care of a physical machine. It also extends the scope of services provided with a printing machine. Additionally, to provide the best services and offer as low downtime as possible, the service providers seek for optimal maintenance strategies. To that end, domain experts need data-driven support to take care of Multifunctional Printers (MFPs). Therefore we analyze the process of maintenance in this domain (section 3) and to support this activity we propose a generic approach to feature extraction for signals from MFPs (section 3.1). Such features give structured information of behavior of machines. They can be utilized by decision support systems or domain experts. To our knowledge, this is the first paper describing feature modeling for the domain of MFPs.

The dataset analyzed in this paper is constructed based on signals sent by the device. It has over 10 000 000 rows and tens of columns. The general statistics show 8.80% of values are missing and a significant portion of rows are incomplete (see section 4). The missing values cause problems when a training dataset for machine learning is being prepared. One solution is an interpolation of data, however, linear interpolation of values is poor for sensor data generated by printers (section 4.2). Having this issue we used linear interpolation as a naïve baseline and with studying common constraints of sensor data we selected and evaluated six methods of interpolation (section 4.3). Further, the results are analyzed and discussed in section 5.

## 2 RELATED WORK

In a study [10], the authors underline the need for structured data coming from devices of Internet of Things (IoT) through meaningful abstraction of the raw input. The motivation for building information abstraction and knowledge representation on the top of IoT data is to reduce the size of data describing IoT devices. This can be beneficial for network transfer and storage. They point out that the data abstraction can be employed as a fundamental base for reducing complexity and network traffic of existing machine learning techniques. The study depicts examples of such approaches in domains of nature, automotive, healthcare, social life.

Christ et al. [6] designed and implemented a software library FRESH [1], that is able to generate generic features for time-series data. The proposed over 100 generic features, like peaks, autocorrelations, Fourier Transform, etc. are related to repetitive time-series. They point out that feature extraction is a crucial part of the machine learning process, and therefore they examine their solution on multiple classification problems. They assume the raw data is clean and missing values are handled.

Another approach to feature extraction for time-series data is presented in [13]. They propose the lower bounding symbolic approach, that allows a numerical time series of arbitrary length to be reduced to a string of arbitrary length. With this feature, measuring distance between two symbolic strings is easier than the calculation of the distance between two time-series. For example, the Jaccard coefficient [12] is only well defined for discrete data as thus cannot be used with real-valued time series.

Yang et al. [21] focus on data imputation for the specific problem of a time-series spatial data. They use characteristics of co-location and spatio-temporal Hidden Markov Model to calculate the fastest route in traffic. Spatial data is out of the scope of our research, so here we would like to stress the development of imputation methods for specific types of data and industry, like logistics in this example.

All cited articles do not describe missing values handling, moreover, some of them even take missing values handling as an assumption and input data constraint.

Mitigation of the problem of missing values is proposed in [8]. The authors describe the preprocessing step of data filling. The first step is feature selection, where the authors propose a distance measure between two features sets and apply to raw data with missing values. The approach is general and does not consider specific aspects of time-series, like a sequence of signals.

*The corresponding author.
First International Workshop on Data Science for Industry 4.0.


Published in the Workshop Proceedings of the EDBT/ICDT 2019 Joint Conference (March 26, 2019, Lisbon, Portugal) on CEUR-WS.org.

However, such methods can also be applied to time-series. Review of general methods and algorithms for preprocessing of sequential data are presented in [15].

# 3 PROCESS OF MAINTENANCE OF MULTIFUNCTIONAL PRINTERS

The MFP is a complex electro-mechanical device consisting of several autonomously working parts [14, 16]. Each part is equipped with its own set of sensors and controllers. These controllers are responsible for sending gathered information from sensors to the centralized database. Each part of the device is able to run self-diagnostics and evaluate its health status. Corrective actions can be initiated automatically when they occur and if the device is not able to return to the functioning state, the information about the error is reported to the centralized database and device itself is set to out of order state.

The MFP's problems collected in the centralized database are distributed to the responsible service departments. Each problem is assigned to a customer engineer who can analyze collected measurements from a device with maintenance need. Further, the engineer can remotely connect to the device when it is supported and collect more diagnostic up-to-date data, and optionally also remotely repair the device. If the remote repair is not possible, then the engineer is dispatched to the customer and solves the issue on site. During the first visit, the engineer verifies the root cause of the problem identified remotely and if needed the necessary spare parts are ordered. With the next visits, the problem is finally solved.

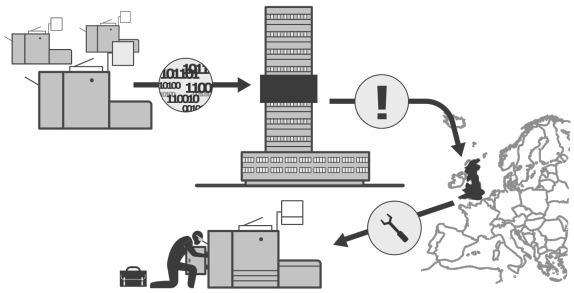

**Figure 1: Process of maintenance of MFPs.**

In the past, the organization, processes, and information systems grew organically which caused a few problems in the whole maintenance process chain. Moreover, the central database storing device measurements are not designed for analytical purposes. The major flaw of the system is missing information about customer engineer intervention- this information can be inferred by observing usage counters of particular parts and looking for sudden drops, i.e., part exchanges. In order to save data transfer, the measured signals are aggregated on the device side. That means the diagnostic data is usually sent to the centralized database on a daily basis. Thus, the normal time-granularity for logged information is one day but, on the one hand, there could be a situation when the device sends information twice a day (e.g., by customer engineer remote request) and, on the other hand, the device does not send information at all for a few consecutive days. The following subsections describe particular device measurements, introduce abstraction used for feature engineering process, and discuss other aspects of data and processes in the real world deployment which must be considered.

## 3.1 Counters and States

MFPs produce a number of signals about the current state of a machine. Definition of the transformation of each type of signal to extract features for Machine Learning models would be exhaustive. We need to define abstract categories of signals where we can assign transformations that prepare data as Machine Learning features.

We realized that each signal can be classified as

- *State*: version of firmware, temperature, humidity, etc.
- *Counter*: number of printed (black/ color) pages, copied pages, number of use of particular part of the device, number of paper jams, etc.

This categorization reduces an effort of the invention of features to only two types of data. The *Counter* type is characterized by a natural number, non-decreasing monotonicity, starting from zero. Also, the *State* can be either number (e.g., temperature) or category (e.g., software version). There is no strict dependency between the previous and the current state (e.g., monotonicity of consecutive feature values), however, for some states, we can define a transition graph describing allowed transitions between states.

## 3.2 Global aspects of data and processes

We used data from multiple divisions (countries and continents) for the feature modeling. We included the name of the division in the dataset, as a feature, that is one of the states of a device.

Multiple obstacles can be faced by following this approach. The first is the integration of data from various types of systems for the logging of device behavior and maintenance, used in each division. The second is the range of temperature and humidity, that is specific for each geographical region. There are at least three possible approaches for such specific data in the context of Machine Learning:

(1) Get data as is, having better coverage of feature space, but less similarity of examples between countries;
(2) Normalize data to $[0, 1]$ range according to the yearly temperatures in each region;
(3) Exclude that information from the dataset to keep similarity between examples over divisions but lost some information on the state of a machine.

We decided to use the data as is, to limit the scope of the described experiment. The second and the third options are transformations, so they will be considered in a separate experiment. In this case, the model is able to decide either to use temperature and humidity for each division separately (using the derived name of division from the features) or globally or not at all.

The third aspect of using global data for a single Machine Learning model is the fact of different rules of maintenance and utilization of MFPs for each division. Rules depend among others on a number of available engineers in a division, cost of maintenance or Service Level Agreements (SLAs). This aspect is also not in the scope of this research.

# 4 EXPERIMENTS

## 4.1 Dataset description

We worked on a dataset that is over several years history of reporting signals by MFPs at Konica Minolta. We selected one model type of MFP because the database was designed for operations, not analytics and the meaning of columns can differ

between different MFP model types. The dataset contains signals we categorize as either *States* or *Counters* according to our observation in the previous section.

The dataset is composed with signals sent by the device. Devices should send at least one signal per day. However, sometimes there is no signal received by the server from a device for several days in a row. It can be caused by, e.g., network problems, or power off of the machine. In consequence, the dataset that contains more than 10 000 000 rows has 8.80% of missing values but from the perspective of incomplete rows, there is a significant portion of rows with at least one missing feature.

## 4.2 Linearity of counter features

In this experiment, we calculated the linear regression model, based on timestamps and values of a feature for each feature and machine (MFP) in the dataset. Then we applied these models to the data to check how linear data is in each feature. First, we aggregated results of predictions per machine and, for each of them, calculate Root Mean Square Error (RMSE) as an appropriate measure of error of models, where the distribution of error is more likely to be Gaussian than the uniform distribution [5]. Then we calculated the second aggregation to 1st quantile (Q1), median, 3rd quantile (Q3) and average. We omit minimum and maximum as prone to be outliers. The results are presented in the table 1.

**Table 1: Linearity of features - RMSE of models for each device**

| metric | mean | Q1 | median | Q3 |
|---|---|---|---|---|
| Jams: | | | | |
| Total | 20239.32 | 7219.14 | 14362.14 | 26411.54 |
| Paper Feed Conveyance | 0.78 | 0 | 0 | 0.42 |
| Automatic Doc. Feeder | 3.51 | 0.01 | 0.73 | 2.38 |
| Tray1 Paper Feed | 2.99 | 0.01 | 0.64 | 2.16 |
| Tray2 Paper Feed | 0.48 | 0 | 0 | 0.38 |
| Print | 15641.15 | 4880.76 | 10520.74 | 20543.13 |
| Parts: | | | | |
| Total | 16479.86 | 5812.15 | 11658.23 | 21411.40 |
| Tray1 Feed Roller | 9851.18 | 1526.41 | 5539.23 | 12619.08 |
| Tray2 Feed Roller | 2631.54 | 113.43 | 404.86 | 1657.03 |
| Drum Rotation Time | 1906.54 | 603.26 | 1239.28 | 2305.48 |

There are two groups of features. The first is related to jams of paper in printers. The characteristic of counters of jams in feeders (Paper Feed Conveyance, Automatic Document Feeder, Tray1/Tray2 Paper Feed) is near linear ($RMSE =< 0.1$) for the first quantile of machines. Counters of the total jams and the print are non-linear (because of large $RMSE > 1000$) even for the first quantile of machines. The second group is related to parts. There is no relation of linearity among counters in this group.

## 4.3 Interpolations

The experiment for feature interpolation is based on a comparison of multiple spline methods. Splines are widely used to fit a smooth continuous function through discrete data. The interpolation algorithm is based on piecewise polynomial fitting. This allows us to use low-order polynomials and reduce computational complexity and numerical instabilities that arise with higher degree curves [19]. Cubic Hermite spline is one of the spline representatives; it is used for interpolation, i.e., line curve passes through the guiding points. Each piece of cubic Hermite spline is determined by values and derivatives (tangents) of its endpoints as is illustrated in Figure 2.

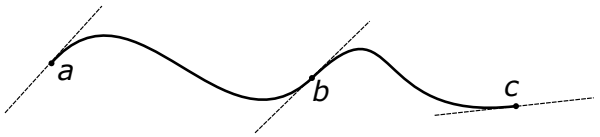


**Figure 2: Spline interpolation on three points a,b,c. Dashed lines show tangents in their respective points.**

Examined methods are $fmm$, that is the spline by [9] (an exact cubic is fitted through the four points at each end of the data, and this is used to determine the end conditions). Method *hyman* computes a monotone cubic spline using Hyman filtering of a method $fmm$ fit for strictly monotonic inputs [11]. The method *monoH* computes a monotone Hermite spline according to the method of [4]. Natural splines [20] and periodic splines [17] are also used in this experiment. For the purpose of the experiment we randomly selected 20% of known data to be missing values in the dataset. Therefore, we can measure the RMSE of the values interpolated by examined methods and real values of features.

**Table 2: Relative RMSE of interpolation. Linear model (lm) as the base: 1**

| metric | fmm | hyman | lm | monoH | natural | periodic |
|---|---|---|---|---|---|---|
| Jams: | | | | | | |
| Total | 536.15 | 608.92 | 1 | 0.07 | 0.08 | 6.67 |
| Doc. Feeder | 0.02 | 0.59 | 1 | 0.01 | 0.01 | 1.82 |
| Tray1 Feed | 0.01 | 0.004 | 1 | 0.004 | 0.01 | 0.30 |
| Tray2 Feed | 0.03 | 0.01 | 1 | 0.01 | 0.01 | 0.39 |
| Control | 67.31 | 0.01 | 1 | 0.01 | 0.01 | 0.76 |
| Print | 12.11 | 94.58 | 1 | 0.07 | 0.08 | 5.92 |
| Parts: | | | | | | |
| Total | 110.14 | 28.37 | 1 | 0.29 | 0.29 | 5.25 |
| Tray1 Roller | 1513.01 | 115.65 | 1 | 0.09 | 0.10 | 13.60 |
| Tray2 Roller | 912.90 | 126.66 | 1 | 0.01 | 0.01 | 4.16 |
| Drum Rot. | 3733.57 | 234.90 | 1 | 0.11 | 0.11 | 14.64 |

The RMSE of *MonoH* is maximally 29% of the linear interpolation and minimally 0.4%. The second top method is a *natural* algorithm. It is the same results for most of the examined metrics. It's worse than *MonoH* for 4 of 10 metrics. The other algorithms are significantly worse than *MonoH* for most of metrics. The worst interpolation method is $fmm$, that is worse than linear interpolation for 7 of 10 metrics.

Figure 3 shows a *Counter* signal of a single arbitrary selected machine. Further, six interpolation methods applied to the raw data are depicted in this figure. We can see the proximity of interpolations to the raw data. *Hyman* and *monoH* estimates real values better than *natural* and $fmm$ methods for interpolations between April and May on the chart. Interpolations of missing values in the second half of April shows that *periodic*, *natural* and $fmm$ does not fit the requirement of monotonicity of the *Counter* value and causing bigger error than *monoH*.

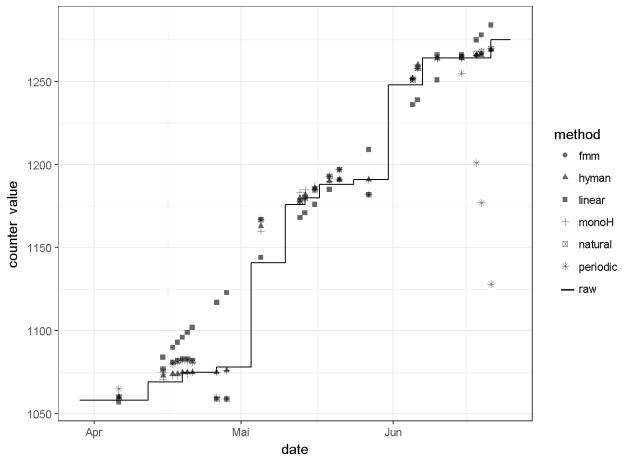


**Figure 3: Example of interpolation of a *Counter* for a single arbitrary selected machine. Raw data as the line.**

## 5 DISCUSSION OF RESULTS

We can see that linearity of features depends on the type of feature even among counters. More general features like Total Jams or Total Parts utilization has a higher error (for all kind of aggregations) than more specific features in each category (i.e., in jams and parts respectively). It is because the total counter is a sum of all counters in a category. Total counters have larger "steps" in its values. It is depicted in figure 4. It is also visible that linear interpolation of a counter with large steps (a) has a higher error than a counter with small steps (b). The largest error is for the sum of them (c). Therefore linear interpolation is a naïve approach to data imputation of MFP counters.

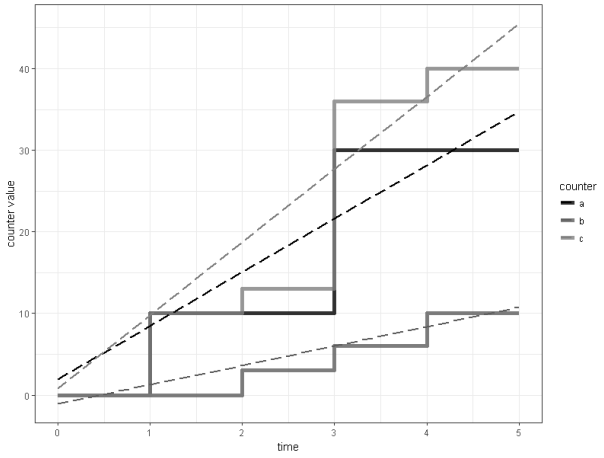


**Figure 4: Example of a linear interpolation of counters (a, b) and sum of them (c). RMSE are 4.61, 1.03, 4.92 respectively**

Linear interpolation works well only for some of the machines and specific signals, like a number of jams for feeders. For those types of signals, the RMSE of the first quantile of results is =< 0.01, which is acceptable from the business perspective. However, the other types of counters, like total jams, print jams, and all examined part counters have non-linear characteristics even for the first quantile of machines. Therefore some more sophisticated methods of interpolation were examined.

Two out of five alternative examined methods of interpolation are better than the linear model. *MonoH* is the best interpolation method for all examined signals. Metrics related to Parts were harder to interpolate than Jams. Jams were easy to interpolate for all methods (relatively to the other metrics).

*MonoH* is significantly better (at least a few times less RMSE) than linear interpolation for both near-linear and non-linear signals. It means the *MonoH* can be a general method of interpolation of missing data for counters from MFPs.

*Linear* interpolation does not fit into sudden changes of counter values. *MonoH* and *hyman* have the smallest RMSE among examined method, significantly lower than the others. Characteristics of the two methods allow accommodating a sudden change of counter with keeping monotonicity of predicted values.

The *periodic* method is not able to fit into the sudden drops of counters, because they are not regular events. In consequence, the predicted regular drops of counters cause large RMSE for that method. An example of this situation is presented on the Figure 3, where the three last values of *periodic* method are the most distant from the real value among all examined methods.

## 6 CONCLUSIONS AND FURTHER WORK

Missing data handling is an essential predicament of time series modeling for MFP generated signals. Described characteristics of features related to such devices helps to select the optimal method of filling of missing values. Conducted experiments confirmed a good fit of the monotone Hermite spline for this specific domain.

This paper is the very beginning stage of defining standards of data processing in the domain of MFPs. We described specific aspects of data of MFP source and propose an approach to generic feature engineering. We selected and examined six approaches to the filling of missing data. Further, we measured the error of each selected method to the ground truth. Based on the findings, we can recommend using *MonoH* and *hyman* interpolation methods in the context of MFP signals.

Our further research will focus on applications of the proposed framework of feature modeling to the input of Machine Learning algorithms for problems relevant to MFPs, like decision support for maintenance of devices, so-called predictive maintenance.

## 7 ACKNOWLEDGMENTS

The authors would like to thank Matej Dusik, Markus Maresch, Dragan Spasic, Arame Shanazari for their invaluable support. This research was supported by Konica Minolta Laboratory Europe.

## REFERENCES

[1] 2018. Automatic extraction of relevant features from time series:: blue-yonder/tsfresh. https://github.com/blue-yonder/tsfresh original-date: 2016-10-26T11:29:17Z.

[2] Zahir Ahamed, Takehiro Inohara, and Akira Kamoshida. 2013. The Servitization of Manufacturing: An Empirical Case Study of IBM Corporation. *International Journal of Business Administration* 4, 2 (March 2013). https://doi.org/10.5430/ijba.v4n2p18

[3] Nils-Petter Augustsson, Jonny Holmstrom, and Agneta Nilsson. 2015. From Technological Transitions to Service Transitions : A Study of Attenuation Effects in IT Service Provisioning. *Journal of the Korea society of IT services* 14, 2 (June 2015), 337–354. https://doi.org/10.9716/KITS.2015.14.2.337

[4] R. Carlson and F. Fritsch. 1985. Monotone Piecewise Bicubic Interpolation. *SIAM J. Numer. Anal.* 22, 2 (April 1985), 386–400. https://doi.org/10.1137/0722023

[5] T. Chai and R. R. Draxler. 2014. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7, 3 (June 2014), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[6] Maximilian Christ, Nils Braun, Julius Neuffer, and Andreas W. Kempa-Liehr. 2018. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* 307 (Sept. 2018), 72–77. https://doi.org/10.1016/j.neucom.2018.03.067

[7] M.A. Cusumano. 2008. The Changing Software Business: Moving from Products to Services. *Computer* 41, 1 (Jan. 2008), 20–27. https://doi.org/10.1109/MC.2008.29

[8] Gauthier Doquire and Michel Verleysen. 2012. Feature selection with missing data using mutual information estimators. *Neurocomputing* 90 (Aug. 2012), 3–11. https://doi.org/10.1016/j.neucom.2012.02.031

[9] George Elmer Forsythe, Michael A. Malcolm, and Cleve B. Moler. 1977. *Computer methods for mathematical computations.* Prentice-Hall, Englewood Cliffs, NJ.

[10] Frieder Ganz, Daniel Puschmann, Payam Barnaghi, and Francois Carrez. 2015. A Practical Evaluation of Information Processing and Abstraction Techniques for the Internet of Things. *IEEE Internet of Things Journal* 2, 4 (Aug. 2015), 340–354. https://doi.org/10.1109/JIOT.2015.2411227

[11] J. Hyman. 1983. Accurate Monotonicity Preserving Cubic Interpolation. *SIAM J. Sci. Statist. Comput.* 4, 4 (Dec. 1983), 645–654. https://doi.org/10.1137/0904045

[12] Michael Levandowsky and David Winter. 1971. Distance between Sets. *Nature* 234 (Nov. 1971), 34. http://dx.doi.org/10.1038/234034a0

[13] Jessica Lin, Eamonn Keogh, Li Wei, and Stefano Lonardi. 2007. Experiencing SAX: a novel symbolic representation of time series. *Data Mining and Knowledge Discovery* 15, 2 (Oct. 2007), 107–144. https://doi.org/10.1007/s10618-007-0064-z

[14] Masakazu Nagano, Yu Iritani, Satoshi Murakami, Thomas Keen, Christoph Gredler, Florent Cuchet, Peter Li, Tom Judd, and Eriko Matsumura. 2015. Electronic copying machine. https://patents.google.com/patent/USD745087S1/en

[15] Sergio Ramírez-Gallego, Bartosz Krawczyk, Salvador García, Michał Woźniak, and Francisco Herrera. 2017. A survey on data preprocessing for data stream mining: Current status and future directions. *Neurocomputing* 239 (May 2017), 39–57. https://doi.org/10.1016/j.neucom.2017.01.078

[16] Ayumi Uchikawa, Masakazu Nagano, and Takashi Terasaka. 2016. Electronic copying machine. https://patents.google.com/patent/USD752134S1/en

[17] Grace Wahba. 1975. Smoothing noisy data with spline functions. *Numer. Math.* 24, 5 (Oct. 1975), 383–393. https://doi.org/10.1007/BF01437407

[18] Jörg Weking, Maria Stöcker, Marek Kowalkiewicz, Markus Böhm, and Helmut Krcmar. 2018. Archetypes for Industry 4.0 Business Model Innovations. *24th Americas Conference on Information Systems (AMCIS 2018)* (Aug. 2018), 11.

[19] George Wolberg and Itzik Alfy. 2002. An energy-minimization framework for monotonic cubic spline interpolation. *J. Comput. Appl. Math.* 143, 2 (June 2002), 145–188. https://doi.org/10.1016/S0377-0427(01)00506-4

[20] Graeme A. Wood and Les S. Jennings. 1979. On the use of spline functions for data smoothing. *Journal of Biomechanics* 12, 6 (Jan. 1979), 477–479. https://doi.org/10.1016/0021-9290(79)90033-2

[21] Bin Yang, Chenjuan Guo, and Christian S. Jensen. 2013. Travel Cost Inference from Sparse, Spatio Temporally Correlated Time Series Using Markov Models. *Proc. VLDB Endow.* 6, 9 (July 2013), 769–780. https://doi.org/10.14778/2536360.2536375