Imputation of Missing Price Values: Supporting Cash-Based Humanitarian Responses in Northern Syria

Pepa Arán Paredes[a,1], Olivier Dietrich[a,1], Mariëlle van Kooten[a,1], and Pierre Winter[a,1]

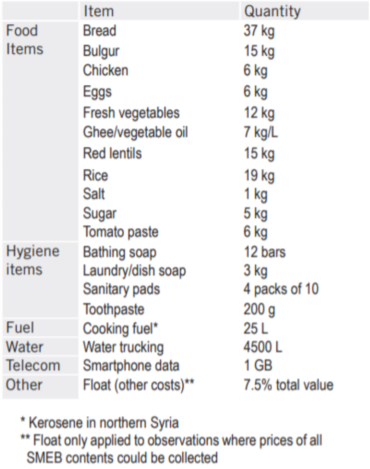
[a]Hack4Good, ETH Zürich, Switzerland; [1]Authors listed in alphabetical order as they contributed equally

Humanitarian aid in the form of cash-based assistance is becoming increasingly popular in regions of conflict (The Cash Learning Partnership (2018)). IMPACT Initiatives provides data-driven solutions to better inform humanitarian cash programming in Syria. This assistance relies on the calculation of a Survival Minimum Expenditure Basket (SMEB), corresponding to the minimum amount of cash necessary to purchase items required to support a 6-person household for one month in a specific geographical area. Price data for these items are acquired through informants in the regions of interest. As a result of temporal inaccessibility due to, most notably, conflict resurgence, prices for certain items may be unattainable and a SMEB cannot be calculated, which makes cash-based assistance challenging. Here, we present analytical methods to impute missing price values. We apply a two-tier strategy in which we show fine-grained imputation to produce a low-error output in data sets where the absence of data points is random, and coarse-grained imputation for data sets where data is missing for specific points in time or for a certain geographical area. These methods allow for the calculation of a SMEB and in effect guarantee a data-driven supply of cash-based aid to Northern Syria and to conflict regions in general.

Since 2011, major conflicts in Syria have led to severe destruction, large-scale displacement, and an escalating humanitarian crisis (IMPACT Initiatives, 2019). Due to the scale and longevity of the conflicts, approximately 13 million people in Syria are now in need of assistance. Many aid programs and non-governmental organizations (NGO's) are currently aiming to provide this needs-based assistance. In order to understand how much assistance is needed, however, the acquisition of timely and comparable information is crucial. A standard way to quantify material aid is by calculating a Survival Minimum Expenditure Basket (SMEB) which represents the minimum, culturally adjusted items required to support a 6-person household for one month.

IMPACT Initiatives aims to improve the impact of humanitarian, stabilisation, and development aid to crisis regions. The Swiss-based NGO provides data-driven solutions to better inform humanitarian cash programming in, amongst others, Syria and monitors market prices to inform partner NGO's of the population's assistance need. IMPACT Initiatives helps provide cash-based assistance to families in northern Syria based on the calculation of a SMEB (Figure 1). While the contents of a country's SMEB may not change often, the prices of the contents will change according to a range of factors, which include normal economic market movements as well as political, military, and environmental factors. It is therefore essential to regularly monitor these markets so that an appropriate amount of cash-based assistance can be delivered to the people in need and to achieve an accurate distribution of assistance across crisis regions. The market research branch of IMPACT Initiatives, REACH, performs price monitoring on a monthly basis in Syria with the use of informants stationed in defined geographical regions. However, given the constantly changing conflict dynamics in regions across Syria, this is often a difficult task to undertake in the field and results in many SMEB item prices to not be documented. Current methods for addressing these missing prices are reported to result in a large over- or under-estimation of the cash-based assistance needed (personal communication within the scope of Hack4Good, IMPACT Initiatives). This has an immediate impact on the population in the conflict area and a long-term impact on the sustainable support from aid organisations.

Here, we assess price data for items acquired in Northern Syria and present and validate methods to impute missing price values. As such, we provide IMPACT Initiatives with a data-driven and benchmarked approach to calculate a SMEB even in the absence of a majority of data points. These methods are broadly applicable to impute price data in conflict areas and can be utilized for data sets that come with a range of sparsities.



	Item	Quantity
Food Items	Bread	37 kg
	Bulgur	15 kg
	Chicken	6 kg
	Eggs	6 kg
	Fresh vegetables	12 kg
	Ghee/vegetable oil	7 kg/L
	Red lentils	15 kg
	Rice	19 kg
	Salt	1 kg
	Sugar	5 kg
	Tomato paste	6 kg
Hygiene items	Bathing soap	12 bars
	Laundry/dish soap	3 kg
	Sanitary pads	4 packs of 10
	Toothpaste	200 g
Fuel	Cooking fuel*	25 L
Water	Water trucking	4500 L
Telecom	Smartphone data	1 GB
Other	Float (other costs)**	7.5% total value

* Kerosene in northern Syria
** Float only applied to observations where prices of all SMEB contents could be collected

Fig. 1. Items in the Syrian Survival Minimum Expenditure Basket (SMEB) (IMPACT Initiatives, 2019).

1. Methods

IMPACT Initiatives has acquired price data for items in a total of 83 unique geographical locations in Syria, referred to

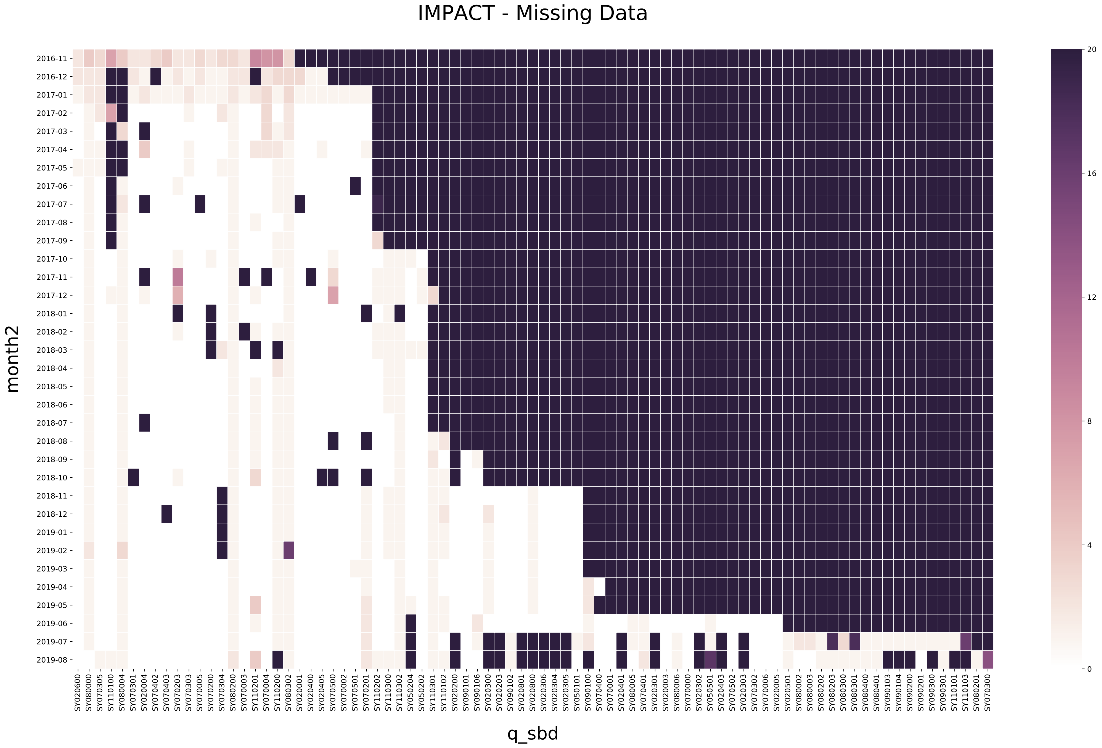


Fig. 2. An overview of *Dataset1* and its distribution of missing values. A maximum of 20 SMEB item prices per total of 83 unique geographical locations in Syria, referred to as subdistricts, for the 34 months spanning November 2016 to August 2019 is available.

as subdistricts, for the 34 months spanning November 2016 to August 2019. Of these items, 20 are necessary for the calculation of a SMEB. This results in an initial data set with (83x34)x20=56,440 potential data points in the form of prices in Syrian pounds (SYP) and it will be called *Dataset1* (Figure 2 and Table 1).

A Complete Dataset to Establish Performance Metrics. To probe a statistical model, subdistricts for which pricing information was unavailable for all 20 objects were removed from *Dataset1*. The resulting data was tabulated into *Dataset2* which contains 1192x20=23,840 data points. The rows in *Dataset2* now have at least 1 out of the 20 SMEB item prices. We then created a training data set which contains no N/A values at all. This allows us to have a baseline which we can use as a ground truth for model comparisons in the future. To create this training data set, which we call *Dataset3*, we removed any row from *Dataset2* which had an N/A value. Since 33.7% of the rows in *Dataset2* contain at least one N/A value, the resulting *Dataset3* is made up of 790x22=15,800 data points. *Dataset3* is a multi-column training dataset and will be used as the ground truth during validation of the proposed imputation methods.

Method Validation. In order to validate our imputation methods, multiple testing sets were created from *Dataset3* using a bootstrapping approach. This consists of removing a predefined percentage of the data at random from *Dataset3*, allowing us to test different levels of "missingness" for a given data set. For a given percentage of missing data, this bootstrapping approach was performed 20 times in order to create 20 different testing sets. The imputation methods were then applied to each of the 20 testing sets. For each run, the error was defined as the normalized root mean squared error (nRMSE):

Table 1. Description of Datasets

	Rows	Columns	# of points	% rows with N/A	% N/A total
Dataset1	2,822	20	56,440	72.0	57.7
Dataset2	1,192	20	23,840	33.7	3.2
Dataset3	790	20	15,800	0	0

$$\text{nRMSE} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_{pred} - y_{true})^2}}{\frac{1}{N}\sum_{i=1}^{N} y_{true}} \qquad [1]$$

where N is the number of imputed datapoints, y_{pred} is the predicted price of the imputed item, and y_{true} is the true price of the item from *Dataset3*.

Overview of Imputation Methods. The following imputation methods were applied to the validation framework in the same way. Each method was tested on 20 bootstrapped testing sets and the average of the nRMSE was tabulated as a performance metric. A minimum baseline performance was calculated by imputing the data from a randomized collection of numbers ranging from the minimum and maximum values present in the data set. For a bootstrapped data set consisting of 80% NA values, imputation at random resulted in an nRMSE of 597%.

Sequential Forward Fill. The function ffill in the Python library pandas was used to sequentially impute NA values. Progressive imputation was performed for a specific object and followed values back in time in a specific subdistrict, over all subdistricts in a specific month, over all subdistricts in a specific year and, finally, over all subdistricts over all points in time. For the remainder of the NA values, the function bfill was applied.

Adapted K-Nearest Neighbors (KNN). The adapted KNN method imputes missing values by considering a K-number of closest existing neighbors. Time series for prices have been produced for each location and SMEB item. From these series, KNN values are drawn for each missing value by identifying K existing prices in the preceding and following months, effectively making a set of 2K values around the data point to be imputed. The missing value is imputed by computing the mean of these KNN values. If there are not 2K existing values in the time series, the mean will be computed on a smaller number of values. It can happen that there are not K existing data points on one side of the missing data point but more than K existing data points on the other side. The algorithm does not return more than K values from either side of the missing value and is thus considered to be an adapted version of the classic KNN algorithm. In the cases where a time series contains no existing prices, the missing values are imputed with the mean of the existing prices over all locations for that given month and item. We performed imputation with this method for K = 1 - 5, 8 and 10.The best results are obtained with K = 2, implying that values far in time should not be used for accurate imputation.

Multivariate Imputation by Chained Equations (MICE). MICE is an R-based imputation method that creates multiple complete data sets based on the available values for a certain variable and its relation to the remaining variables in the data set (van Buuren and Groothuis-Oudshoorn (2011)). It assumes that the data is missing at random, meaning the probability that a value is missing depends only on observed values and not on unobserved values. The method is designed for cases in which there are values missing for more than one variable. The MICE algorithm first fills in the missing values by taking the mean over SMEB items to produce an initial complete data set. The algorithm then performs sequential imputation by setting the initially missing values back to NA and filling them according to the `method` passed to the `mice()` function in `R`. In this case we used the `norm.predict` method which fits a ridge regression model with the observed values as a response and imputes the missing values with the corresponding predicted values. The final complete data set is pooled from the complete data sets generated after each MICE iteration (Figure 3). The result is an empirical density of prices for each individual SMEB item.

2. Results

We aimed to provide distinct and simple metrics to impute a sparse price matrix that, here, contains price data for 83 subdistricts and 20 objects over a period of 34 months. A total of 57.7% of potential data points were not present in *Dataset1*. Imputation based on time, based on location, and based on grouped objects or a combination of these factors was considered. A data-driven decision on performance of these factors in imputation necessitates the establishment of trends throughout factor data - as such, a predictable change in time (column-wise trend) or a set object proportionality (row-wise trend) would underlie imputation. Considering the large amount of missing values and the necessity of providing simple, readily accessible methods to IMPACT Initiatives, the authors decided on a simple, two-tier approach where fine-grained imputation can be performed in conjunction with coarse-grained imputation where data is sparse. As seen in Figure 4, the R-based method `norm.predict` performs best for our validation scheme defined in Section 1 with an nRMSE of 23-25% depending on the percentage of NA values in the testing set. However, this method utilizes secondary object information, where relationships between objects have not been established. In addition, deviations are observed to be more extensive than for other MICE methods. Our adapted KNN approach performs similarly well with an nRMSE of 25-30% and is observed to carry less of a deviation. Both methods, however, rely on imputation based on data collected at a future time point. We thus chose to establish an additional method based solely on values generated at time points in the past, referred to as the sequential Forward Fill method (ffill). The ffill approach performs slightly worse with an nRMSE of 32-40%. All NA values were imputed in this data set and performance metrics were assessed for each model. Imputation at random performs worse by a factor of approximately 20 when compared to our proposed methods. Because our methods are straightforward, reproducible, and not parameter-dependent, they can be readily applied to any conflict area of interest in order to improve the data underlying cash-based assistance. The scripts underlying this work have been made available open source in Python and R.

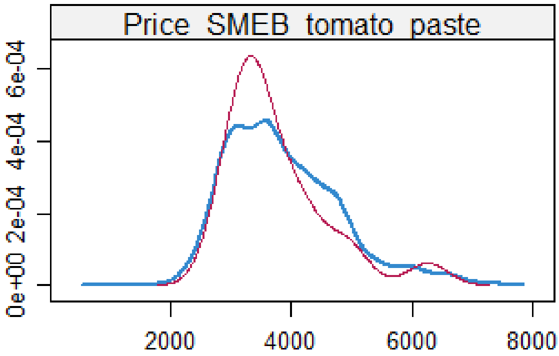


Fig. 3. The empirical density of the observed data (red) and the imputed data (blue) for the object *tomato paste* in *Dataset3* as acquired through the R-based package MICE.

3. Discussion

We made use of chained equations for imputation in a sparse data set that underlies humanitarian cash programming. Of the approaches tested, we show that the R-based MICE pack-

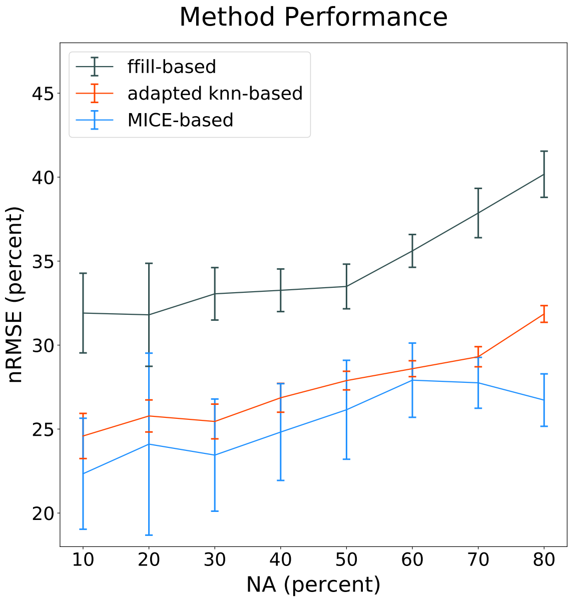


Fig. 4. Normalized root mean square error of imputation methods ffill, KNN and MICE on randomized imputation of *Dataset3*. As would be expected, the normalized RMSE score increases with the percentage of values removed and imputed.

age method `norm.predict` performs best upon randomized removal of price data from a synthetic, complete test data set based on real-world data (Figure 4). Because missing price values are likely not to occur completely at random in a crisis setting, our reported error, which was based on randomized removal of validation data, is likely to underpredict the true error for all models. In addition, imputation with MICE and KNN makes use of price values that include future prices and is thus only possible in an already acquired data set. Finally, the basis of imputation lies in application of recurrent trends to impute missing values and thus on non-random behaviour of price data in conflict areas. In this regard, uncovering hidden factors influencing price data as well as inter-object connections that move in a concerted manner will allow us to impute based on trends seen in partnering objects. These influential factors also would allow for more accurate and computationally expensive methods to be implemented by IMPACT Initiatives. In future explorations of the data sets linked to this initiative, it would be of particular interest to extrapolate local data to new situations. In addition, we recommend building a season-based universal survival minimum expenditure basket (uSMEB) across conflict regions to secure quick implementation of cash programming in arising conflict areas. A uSMEB would carry price data for objects in set relative fractions - in essence, a single price can be used to calculate the value of a complete basket. Taken together, the crucial balance of information need and the rapid supply of cash-based humanitarian aid can be solidified in multiple ways that allow for follow-up assessment.

ACKNOWLEDGMENTS. We are thankful to have been given the opportunity to team up with IMPACT Initiatives. We would like to acknowledge team mentor Mario Tomasello, who has pointed out options as well as pitfalls. The Hack4Good organizing team and Student Project House have provided us with a great working environment and have enabled inter-team communication, which we are certain has been beneficial to more than our team alone.

References

IMPACT Initiatives (2019). IMPACT Syria.

The Cash Learning Partnership (2018). The State of the World's Cash Report (SWCR). page 3.

van Buuren, S. and Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software, Articles*, 45(3):1–67.