Official Journal Issued by Faculty of Veterinary Medicine

Benha Veterinary Medical Journal

Journal homepage: https://bvmj.journals.ekb.eg/

Since 1990

Original Paper

Comparative study of single imputation techniques for the prediction of missing dairy data

Ahmed A. Ahmed[1,*], Eman A. Manaa[2], Basant M. Shafik[2], Saqr A.Mustafa.[3], Ahmed M. Gad[4,5]

[1]*Department of Animal Wealth Development (Biostatistics),* [2]*Department of Animal Wealth Development (Animal and Poultry Production), Faculty of Veterinary Medicine, Benha University, Moshtohor, Toukh, 13736, Qalyubia, Egypt.*

[3]*Animal Production Research Institute, Agricultural Research Center, Ministry of Agriculture, Dokki, Giza, 12619, Egypt..*

[4]*Statistics Department, Faculty of Economics and Political Science, Cairo University, Giza, Egypt.*

[5]*Business Administration Department, Faculty of Business Administration, Economics, and Political Science, British University in Egypt, Cairo, Egypt.*

ARTICLE INFO

Keywords

Imputation methods

Missing data

Record keeping

Expectation maximization

Power regression

Received 03/05/2025
Accepted 09/06/2025
Available On-Line 01/07/2025

ABSTRACT

Dairy farm records are a crucial component of effective livestock business management. Record analysis allows a farm's owner to make informed decisions. Incomplete records are less useful for data analysis, so it's important to handle missing values correctly. This study compares different imputation methods for handling missing values in a dataset of dairy records comprising 997 records collected from 234 cows between 2012 and 2022. The dataset was screened against records with missing values and then deleted, resulting in 858 observations from 200 animals. There were missing values in two variables, with a missing percentage of 13.9%: days in milk (DIM) and total milk yield (TOTM). Then, cases with known values that show the same percentages of missing data as the original dataset for DIM and TOTM are randomly excluded. Five different imputation techniques were compared to obtain the best imputation technique. These techniques include mean imputation, median imputation, power regression imputation, multiple regression imputation, and expectation-maximization method (EM). The results showed that the expectation maximization method was the best imputation method for the data under study. It has the lowest mean absolute deviation MAD (37.54), the lowest mean square error MSE (15425.07), the highest Spearman's correlation coefficient (0.967) and the second lowest mean absolute percentage error MAPE (5.27) for predicting the missing data in missing variables. Power regression imputation comes after expectation-maximization (EM) in predicting missing values, as it gives results better than other imputation methods but lower than Expectation-maximization (EM).

1. INTRODUCTION

Record keeping is an essential component of effective livestock business management. It is preserved to help with providing data for government administrative and extension purposes, aiding in livestock management decisions, making financial planning decisions, and analyzing overall dairy farm activities (Grisham, 2007). So, dairy farm records can boost farm profitability by tracking from birth to death and acting as a source of knowledge later. Without keeping accurate records and avoiding uncertainty, it is impossible to make wise decisions on farms (Patil and Patil, 2020). Large volumes of data can be recorded thanks to the deployment of dairy records. The gathered data are crucial to the production of animals because they may yield valuable information that improves forecasts, plans, and data-based decisions that give animals more benefits from a welfare, feeding, or healthcare perspective and increase the sustainability of animal production on an economic, environmental, and social level (Menendez et al., 2022).

Ideally, there would never be any missing data in your dataset. However, in biology or any other science, excellent datasets are uncommon (Shinichi, 2015). Three general types of missing data mechanisms are classified into: (a) missing completely at random (MCAR), where there is no bias and records are a random sampling of the intended records; (b) missing at random (MAR), where the available variables can account for the missing data (for example, we can statistically account for the bias and are aware of it); and (c) missing not at random (MNAR), where the existing information is insufficient to explain the missing data (for example, we do not know about the bias or do not have relevant data that could statistically explain it) (Johnson et al., 2021). A dataset may contain missing values for a variety of reasons, including clerical errors during data entry, system malfunctions, or data corruption during storage and transmission (Arefn et al., 2024). Since most data analysis techniques rely on complete datasets, most statistical procedures do not include observations with any missing values in their analysis, as the incomplete data is less useful. While using only complete cases is simple, it results in a loss of information regarding the incomplete cases. Additionally, this method disregards the potential systematic variation between complete and incomplete cases, making the resulting inferences less applicable to the overall population, especially when the number of complete cases is less

* Correspondence to: ahmed.abdelhakim@fvtm.bu.edu.eg

(Dettori et al., 2018). It is critical to address the missing value issue with effective techniques that produce a complete dataset. The most popular technique for handling missing data in counseling research is listwise deletion, which is the default setting in SPSS software and other similar software programs like STATA and SAS (Sterner, 2011). According to earlier research, imputation typically performs better in biological studies than complete-case analysis. Imputation, however, is only effective if it takes into consideration the mechanism underlying missing data. It is possible that imputation could even increase inference error if the imputation model is unable to explain this mechanism (for example, when subjected to severe biases like MNAR) (Penone et al., 2014; Kim et al., 2018). Numerous techniques have been established for the data imputation approach, including mean, median, and regression imputation. However, it remains challenging to ascertain the differences between these methods or how they perform when utilized on real-world datasets. Therefore, this study aims to assess and compare the efficacy of several imputation strategies from a commercial dairy dataset. Five univariate and multivariate data imputation techniques were evaluated using dairy cow data that included missing values in two variables (Days in milk (DIM) and Total milk yield (TOTM)).

2. MATERIAL AND METHODS

Ethical approval

The Committee of Animal Care and Welfare at Benha University's Faculty of Veterinary Medicine in Egypt gave its approval to the current study (BUFVTM 45-11-23).

2.1. Data collection

The current study's data included daily records of Holstein dairy cows gathered from a private farm in Nubariyah during the period extending from 2012 to 2022. All animals were housed in open sheds throughout the year, with a cool spraying system in the summer.

Animals were fed on total mixed ration (TMR) all through the year and had free access to water throughout the day. Before and after milking, udder sanitation protocols were followed. Cows were milked twice daily with a milking machine. Cows were estimated in heat by visual inspection and/or excessive activities recorded by the pedometer; they were introduced to insemination 10 to 16 hours later. Manual recording procedures are employed in this dairy farm. Depending on manual data entry as the total milk yield, drying date, and days in milk.

There were 234 cows in the raw dataset, with 997 observations including 139 missing cases representing 13.9% and 858 complete cases representing 86.1% as shown in table 1. And figure 2. The dataset containing 17 variables, including the cow identification number (ID), Sire registration number (SREG), Dam registration number (DREG), Lactation number (LACT), Cow birth date (BDAT), Fresh date (FDAT), Conception Date (CDAT), Age at first calving (AFC), Times to bred (TBRD), Days open (DOPN), Calf weaning Date (WDAT), Calf weaning Weight (WW), Dam Weight (DW, Kg), Calf birth weight (BW), Drying date (DD), Days in milk (DIM), Total milk yield (TOTM). The variables with missing values were Days in milk (DIM) and Total milk yield (TOTM).

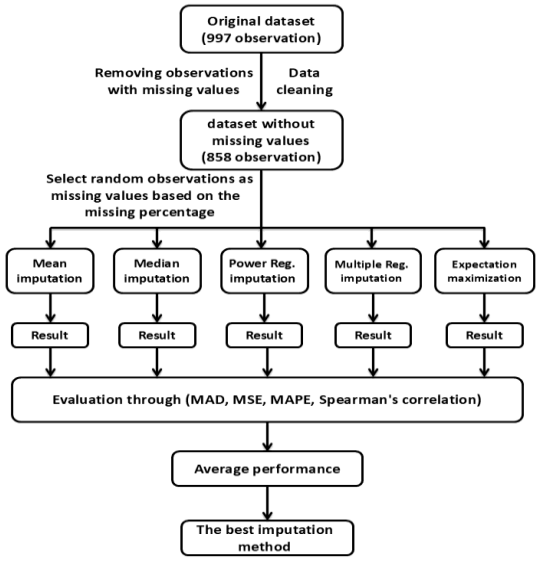

Fig. 1. Workflow diagram of the current study for predicting missing dairy cow DIM and TOTM data. Abbreviations: MAD = Mean Absolute Deviation; MSE= Mean Square Error; MAPE = Mean Absolute Percentage Error.

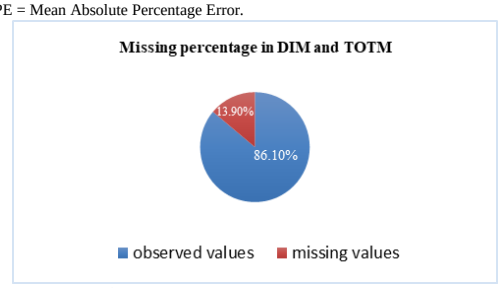

Fig 2 Percentage of Missing values in DIM and TOTM missing variables

Table 1 Show number and percentage of the missing numerical variables.

Missing variables	Valid		Missing		Total	
	N	percent	N	percent	N	percent
DIM	858	86.1%	139	13.9%	997	100%
TOTM	858	86.1%	139	13.9%		

SD: standard deviation, DMY: average daily milk yield, TOTM: total milk this lactation

2.2. Data preprocessing

The raw dataset was cleaned by deletion of missing cases to evaluate different imputation techniques (You *et al.*, 2023). Every set of input variable value has an equivalent known output variable value in order to enable proper testing of imputation results. Following the deletion of all records containing missing values, the original dataset (997 records for 234 cows) was reduced to 858 observations for 200 cows. Observations with known values were randomly removed from this reduced dataset based on the original dataset's percentage of missing values for DIM and TOTM in order to evaluate the effectiveness of imputation techniques in imputed variable values on datasets that display the same structure as the original dataset. Based on the missing percentages in the original dataset, a random selection of values in the DIM and TOTM columns was replaced with missing values. Consequently, two data sets were produced: one for the selected values and one for the missing values. In the first part, missing values were predicted using imputation techniques, and in the second part, the predicted values were compared to the actual values (You *et al.*, 2023). The current study assumed that imputation methods on the cleaned dataset would perform similarly to those on the original dataset (with real missing values). Consequently, the original dataset might be subjected to the best imputation technique based on average performance.

2.3. Imputation methods

2.3.1. Mean imputation

The overall predictor mean derived from the available data for the predictor is used to replace missing values in mean imputation. As a result, a missing value is replaced with the mean of the observed values for the same variable (Moons *et al.*, 2006). This method's drawback is that it frequently fails to adequately account for outliers, which reduces variance. This problem results in a single, constant value being used to replace several missing values (McKnight *et al.*, 2007).

2.3.2. Median imputation

Given that the mean is impacted by the existence of outliers, it would seem reasonable to utilize the median in order to ensure robustness. In this case, the missing data for a given variable is replaced by the median of all known values of that variable (Acuna and Rodriguez, 2004).

2.3.3. Power and multiple regression imputations

Depending on how independent and dependent variables are related, imputation combined with regression on one or more other variables may yield more intelligent results. The variable of interest must be set as the response variable, and other relevant variables can be placed as covariates for researchers to fit a regression model. After estimating the coefficients, the fitted model can be used to predict missing values. For example, a power regression model between DIM and TOTM can be constructed using the dataset under study.

$$y = a.x^b \quad (1)$$

where *y* is the response variable, *x* is the predictor variable; *a and b* represent the regression coefficients that describe the relationship between *x* and *y*.

Another example of multiple regression between (LACT), (AFC), (TBRD), (DOPN), (DW, Kg) as independent variables to predict (DIM) and (TOTM) as missing variables.

2.3.4. Expectation-Maximization (EM)

It is a broad approach for dealing with missing data that preserves the relationship between the missing data and the unknown parameters of the data model. Determining the model parameters is simple when the missing data is known. In a similar way, knowing the parameters makes it easy to estimate the missing values.(Firat *et al.*, 2010). Only datasets where the missing values are missing at random can use the EM algorithm and the techniques that will be derived from it (Schneider, 2001). SPSS can carry out this process by choosing the EM feature in the Missing Value Analysis package. For the partially missing data, the EM (Expectation-Maximization) approach assumes a distribution and relies on conclusions based on the likelihood under that distribution. An E step and an M step make up each repetition. Using the current estimations of the parameters and the observed values, the E step determines the conditional expectation of the "missing" data. Next, these expectations are used instead of "missing" data. The M step calculates maximum likelihood estimates of the parameters as if the missing data were filled in (Kosova and Hajrulla, 2024).

2.4. Model evaluation

The current study used four measures.

2.4.1. The Mean Absolute Deviation (MAD):

The MAD evaluates prediction accuracy by averaging the absolute errors between the forecasted and actual values without considering their direction (positive or negative), which is beneficial for measuring prediction errors in the same units as the original series (Moon and Yao, 2011). A lower MAD value indicates better forecast accuracy. MAD is calculated as

$$MAD = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n} \quad (2)$$

where y_i denotes the actual value of the i[th] sample; $\hat{y}_i$denotes the imputed value of the i[th] sample; n denotes the number of observations.

2.4.2. The Mean Absolute Percentage Error (MAPE):

It provides a clear and understandable indicator of how much predictions deviate from actual values by quantifying the prediction error as a percentage (Templ and Ulmer, 2024). The lower the value for MAPE, the better.

MAPE is calculated as

$$MAPE = \frac{\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i}}{n} \times 100\% \quad (3)$$

where y_i denotes the actual value of the i[th] sample; $\hat{y}_i$ denotes the imputed value of the i[th] sample; n denotes the number of observations.

2.4.3. The Mean Square Error (MSE)

It is obtained by squaring the distances from the points to the regression line (these distances are the "errors"). The lower the MSE, the better the forecast (Hodson et al., 2021).

The MSE is calculated as

$$MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n} \quad (4)$$

where y_i denotes the actual value of the i[th] sample; $\hat{y}_i$denotes the imputed value of the i[th] sample; n denotes the number of observations.

2.4.4. Spearman's rank correlation coefficient (r_s)

Correlation is a crucial factor in feature selection methods to calculate the similarity among the attributes (Shantal et al., 2023). Its value ranges from -1 to +1, depending on the strength of the correlation. A correlation coefficient value near one suggests a strong correlation, whereas a value near zero indicates a weak correlation (Shantal et al., 2023). The Spearman coefficient of rank correlation formula is (Ali and Al-Hameed, 2022):

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)} \quad (5)$$

where d represents the difference in rank between the two variables and n represents the number of observations

All the analytic procedures were performed using the SPSS program version 25

3. RESULTS

The performance metrics for evaluating various imputation techniques for predicting DIM and TOTM missing dairy cow data were shown in Table (2). For the DIM missing variable, the Expectation Maximization method gained the lowest MAD (4.18), followed by power regression imputation (4.39), mean imputation (4.87), and median imputation (4.95), but multiple regression imputation had the highest MAD (6.07).

According to the MSE results, the Expectation-Maximization method showed the lowest MSE (198), followed by power regression imputation (212.7). Mean imputation and median imputation showed relatively the same result (307.8 and 308.7, respectively), and multiple regression imputation had the highest MSE (377.3). The Spearman's correlation coefficient confirms the results of MAD and MSE. The Expectation Maximization method gained the highest r_s (0.936), followed by the power regression imputation (0.927), the Mean imputation and the

Median imputation showed relatively the same result (0.919 and 0.918, respectively). However, multiple regression imputation had the lowest correlation (0.895). The power regression imputation showed the lowest MAPE (2.15), followed by the Expectation Maximization method (2.20). The mean imputation and the median imputation showed relatively the same result (2.69 and 2.80, respectively), and the multiple regression imputation had the highest MAPE (3.19).

Table 2 Model Evaluation of variate imputation methods for dataset under study

Imputing variables	Imputation methods	MAD	MSE	r_s	MAPE
DIM	Mean imputation	4.87	307.8	0.919	2.69
	Median imputation	4.95	308.7	0.918	2.80
	Power regression imputation	4.39	212.7	0.927	2.15
	Multiple regression imputation	6.07	377.3	0.895	3.19
	Expectation maximization	4.18	198	0.936	2.20
TOTM	Mean imputation	64.70	46073.85	0.905	18.12
	Median imputation	64.24	46163.22	0.908	18.60
	Power regression imputation	47.31	23042.45	0948	4.65
	Multiple regression imputation	51.71	30944.66	0.940	5.31
	Expectation maximization	37.54	15425.07	0.967	5.27

Abbreviations: MAD = Mean Absolute Deviation; MSE = Mean Square Error; r_s = Spearman's rank correlation coefficient; MAPE = Mean absolute percentage error; DIM = Days in milk; TOTM = total milk yield.

For the TOTM missing variable, the Expectation Maximization method gained the lowest MAD (37.54), the lowest MSE (15425.07), the highest Spearman's correlation coefficient (0.967), and the second lowest MAPE (5.27) after power regression imputation (4.65). From all the above, the Expectation Maximization method showed the best results of all the performance metrics for predicting missing dairy cow DIM and TOTM data. As shown in Figs. 3 & 4, they show the results of evaluation models for five different imputation techniques in DIM and TOTM missing variables, respectively.

4. DISCUSSION

Maintaining accurate records is critical to running a successful livestock business. It is maintained to support livestock management decisions, provide information for government administrative and extension needs, detect health and welfare obstacles, improve reproductive and productive performance, decrease financial loss and environmental impact, aid in financial planning, and evaluate the general operations of dairy farms (Grisham, 2007). Records should be maintained on the farm. For instance, without precise documentation of deworming, vaccination, timely artificial insemination, pregnancy diagnosis, and cow drying, all of these processes will be compromised, potentially affecting the health of the animals as well as the farm's profitability. During the animals' subsequent care by different vets (Patil and Patil, 2020). A farm records system should provide accurate and essential data (without missing values) that aids decision-making and prediction through analysis of the collected data, as incomplete data can lower the value of the dataset and affect data-driven predictions and decisions (Jeyabalan, 2010). So, using the imputation technique is an essential step before proceeding with downstream data analysis to deal with the missing values problem (Shantal et al., 2023). This study sought to estimate missing DIM and TOTM values in the dataset of a dairy farm by examining several imputation techniques, enabling the data's application in subsequent modeling and analysis. considered as The EM technique has the benefit of producing accurate results when dealing with large datasets or when the missing data are ignorable (Boucher, 2011). Nguyen (2021) also demonstrated that EM is an effective technique to handle missing data, then power regression imputation technique came after the EM technique and also predicted DIM and TOTM better than other imputation methods. Mean and median imputation procedures impute each missing value using the mean or median of the observed values of an incompletely recorded variable, and the results show similar performance with just slight changes. Moreover, they ignore the relationship with other variables. Therefore, when there is a correlation between variables, these approaches perform badly.(Waljee et al., 2013). It would seem reasonable to use the median to provide robustness, as the presence of outliers affects the mean (McKnight *et al.*, 2007).

5. CONCLUSIONS

In comparison to the other imputation techniques, the Expectation-Maximization technique had the lowest MAD, the lowest MSE, the greatest Spearman's correlation coefficient, and the second-lowest MAPE for DIM and TOTM. Overall, this study concluded that the Expectation-Maximization method outperformed all other imputation techniques in predicting DIM and TOTM.

CONFLICT OF INTEREST

The authors announce that they have no conflict of interest.

6. REFERENCES

1. Acuna, E., Rodriguez, C., 2004. The treatment of missing values and its effect on classifier accuracy. Classification, Clustering, and Data Mining Applications. Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, Springer, 15–18 July 2004 pp. 639-647.
2. Ali, A. K., 2022. Spearman's correlation coefficient in statistical analysis. International Journal of Nonlinear Analysis and Applications. 13,1 , 3249-3255.
3. Arefin, M. N., and Masum, A. K. M., 2024. A Probabilistic Approach for Missing Data Imputation. Complexity. 2024, 4737963.
4. Boucher, G., 2011. Book Reviews: Book Reviews. Critical Sociology. pp. 493–497.
5. Dettori, J.R., Norvell, D.C., and Chapman, J.R., 2018. The Sin of Missing Data: Is All Forgiven by Way of Imputation?. Global Spine Journal. 8,8 , 892–894.
6. Firat, M., Dikbas, F., Koç, A. C., and Gungor, M., 2010. Missing data analysis and homogeneity test for Turkish precipitation series. Sadhana Academy Proceedings in

Engineering Sciences. 35,6 , 707-720.
7. Grisham, E., 2007. Record-keeping systems adoption by Louisiana dairy farmers. Thesis, Master of Agricultural & Mechanical College, Louisiana State University, United States.
8. Hodson, T. O., Over, T. M., and Foks, S. S., 2021. Mean squared error, deconstructed. Journal of Advances in Modeling Earth Systems. 13,12 , 1-10.
9. Jeyabalan, V., 2010. Individual cow recording and analysis system for small scale dairy farmers in Malaysia. International Journal of Computer Applications. 8,11 , 33-38.
10. Johnson, T. F., Isaac, N. J., Paviolo, A., and González-Suárez, M., 2021. Handling missing values in trait data. Global Ecology and Biogeography. 30,1 , 51-62.
11. Kim, S. W., Blomberg, S. P., and Pandolfi, J. M., 2018. Transcending data gaps: a framework to reduce inferential errors in ecological analyses. Ecology Letters. 21,8 , 1200-1210.
12. Kosova, R., Naço, A., Hajrulla, S., and Kosova, A. M., 2024. Addressing missing data in surveys and implementing imputation methods with SPSS. International Journal of Advanced Natural Sciences and Engineering Researches. 82, 40-50.
13. McKnight, P.E., McKnight, K. M., Sidani, S., Figueredo, A. J., 2007. Missing data: A gentle introduction. DA Kenny, Ed. Guilford Press, New York, NY, US, 23, pp. 30–54.
14. Jacobs, M., Remus, A., Gaillard, C., Menendez III, H. M., Tedeschi, L. O., Neethirajan, S., Ellis, J. L., 2022. ASAS-NANP Symposium: Mathematical Modeling in Animal Nutrition: Opportunities and challenges of confined and extensive precision livestock production. Journal of Animal Science. 1006, 1–19.
15. Menendez, H.M.; Brennan, J.R.; Gaillard, C.; Ehlert, K.; Quintana, J.; Neethirajan, S.; Remus, A.; Jacobs, M.; Teixeira, I.A.; Turner, B.L.; et al. 2022. ASAS–NANP Symposium: Mathematical Modeling in Animal Nutrition: Opportunities and challenges of confined and extensive precision livestock production. J. Anim. Sci. , 100, skac160. doi: 10.1093/jas/skac160.
16. Moon, Y., and Yao, T., 2011. A robust mean absolute deviation model for portfolio optimization. Computers & Operations Research. 38,9 , 1251–1258.
17. Moons, K. G., Donders, R. A., Stijnen, T., Harrell Jr, F. E., 2006. Using the outcome for imputation of missing predictor values was preferred. Journal of clinical epidemiology. 59,10 , 1092–1101.
18. Nguyen, L., 2021. Handling Missing Data with Expectation Maximization Algorithm. GRD Journal for Engineering, 6,11 , 9–32.
19. Patil, P.V., Patil, M.K., 2020. Dairy Farm Records and Their Maintenance. Milk Production Management. pp. 94–99.
20. Penone, C., Davidson, A. D., Shoemaker, K. T., Di Marco, M., Rondinini, C., Brooks, T. M., Costa, G. C., 2014. Imputation of missing data in life history trait datasets: Which approach performs the best?. Methods in Ecology and Evolution. 59, 961-970.
21. Schneider, T., 2001. Analysis of incomplete climate data: Estimation of Mean Values and covariance matrices and imputation of Missing values. Journal of Climate. 14,5 , 853–871.
22. Shantal, M., Othman, Z., and Bakar, A.A., 2023. Impact of Missing Data on Correlation Coefficient Values: Deletion and Imputation Methods for Data Preparation. Malaysian Journal of Fundamental and Applied Sciences. 19,6 , 1052-1067.
23. Shinichi, N., 2015. Missing data: mechanisms, methods, and messages. Ecological Statistics: Contemporary Theory and Application. pp. 1–53.
24. Sterner, W., 2011. What is missing in counseling research? Reporting missing data. Journal of Counseling and Development. 89,1 , 56–62.
25. Templ, M., and Ulmer, M., 2024. The impact of misclassifications and outliers on imputation methods. Journal of Applied Statistics. 51,14 , 2894-2928.
26. You, J., Ellis, J. L., Adams, S., Sahar, M., Jacobs, M., Tulpan, D., 2023. Comparison of imputation methods for missing production data of dairy cattle. animal, 17, 100921.