US 20250265139A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2025/0265139 A1**

NAONO et al. (43) **Pub. Date: Aug. 21, 2025**

(54) **MISSING DATA IMPUTATION DEVICE AND MISSING DATA IMPUTATION METHOD**

(71) Applicant: **Hitachi, Ltd.**, Tokyo (JP)

(72) Inventors: **Ken NAONO**, Tokyo (JP); **Hiroaki MASUDA**, Tokyo (JP); **Mika TAKATA**, Tokyo (JP); **Tsunehiko BABA**, Tokyo (JP)

(21) Appl. No.: **18/973,505**

(22) Filed: **Dec. 9, 2024**

(30) **Foreign Application Priority Data**

Feb. 16, 2024 (JP) 2024-022296

Publication Classification

(51) **Int. Cl.**
G06F 11/07 (2006.01)

(52) **U.S. Cl.**
CPC ***G06F 11/0772*** (2013.01); ***G06F 11/0727*** (2013.01)

(57) **ABSTRACT**

There are provided an imputation priority flagging unit that gives an imputation priority flag to at least one entry falling within a predetermined proportion defined by an imputation amount adjustment parameter in pieces of data of specific column data items, and a priority order determination unit that counts the number of entries each given the imputation priority flag in entries included in a data table, and that determines an integrated imputation priority order in a descending order of the number of the imputation priority flags.

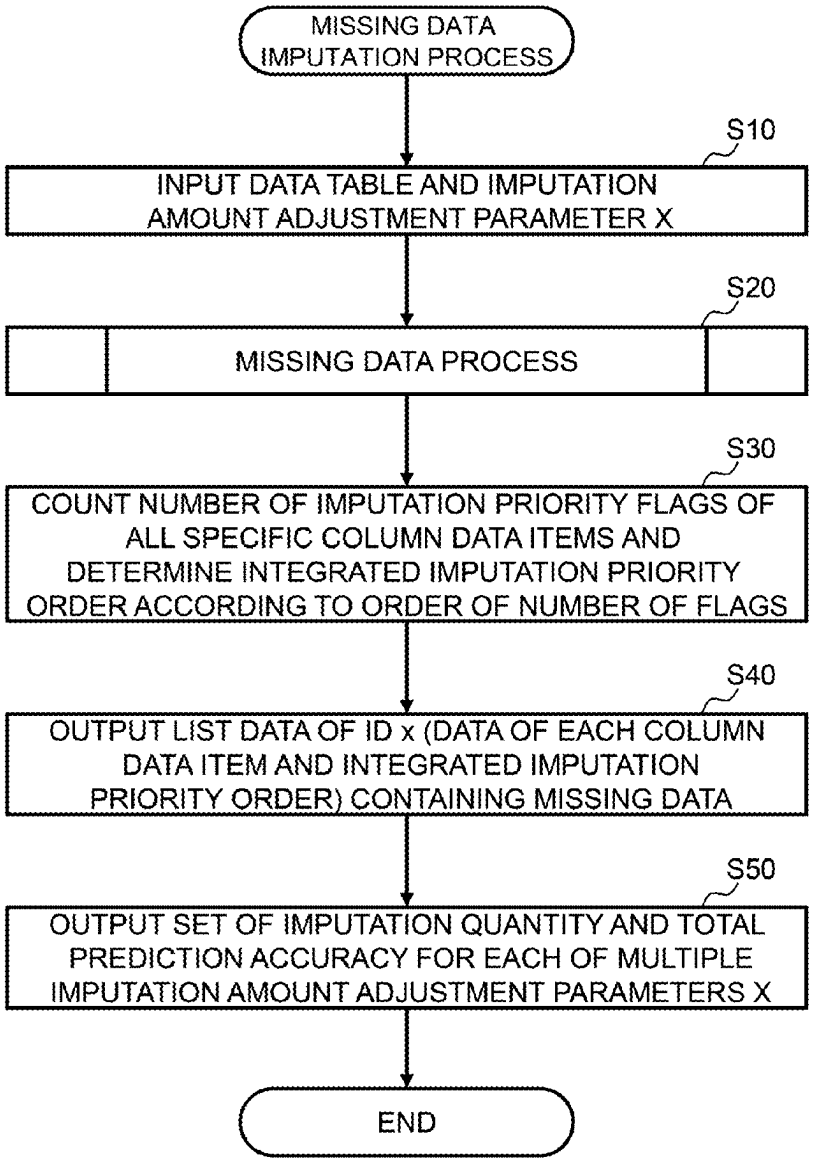


FIG. 1

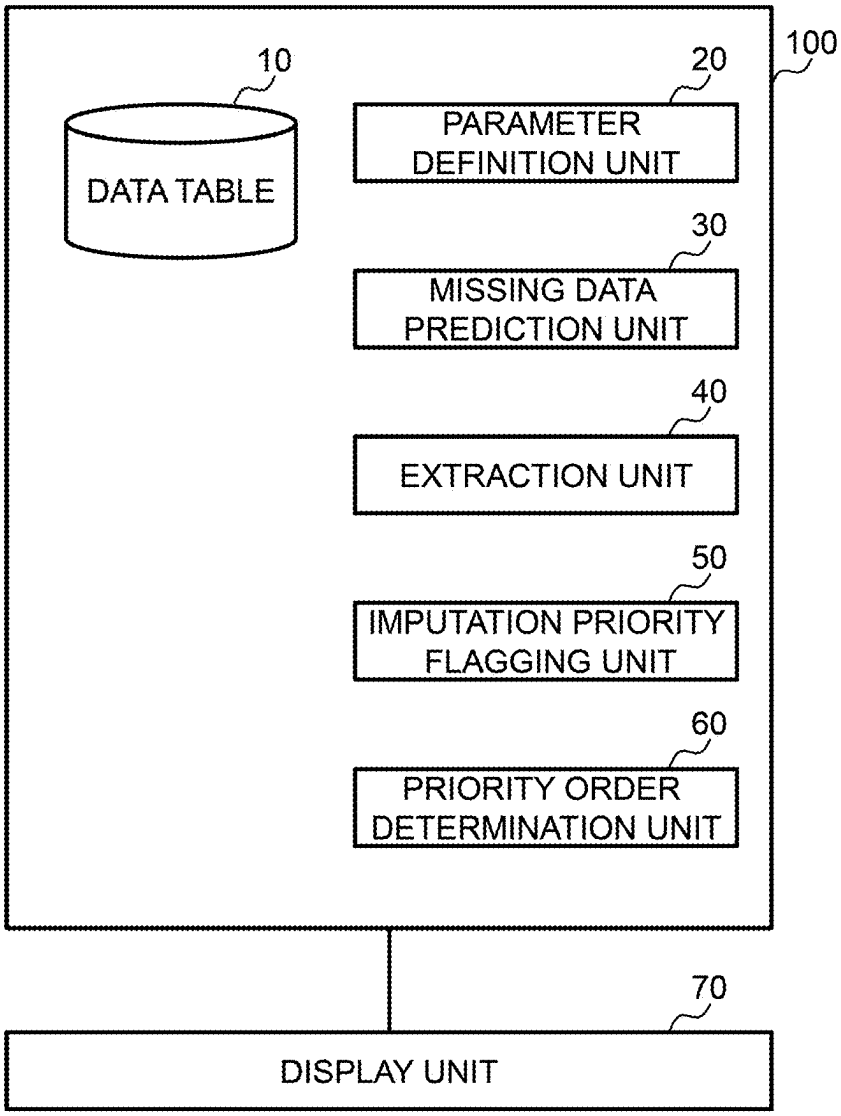

100
10
DATA TABLE
20
PARAMETER DEFINITION UNIT
30
MISSING DATA PREDICTION UNIT
40
EXTRACTION UNIT
50
IMPUTATION PRIORITY FLAGGING UNIT
60
PRIORITY ORDER DETERMINATION UNIT
70
DISPLAY UNIT

FIG. 2

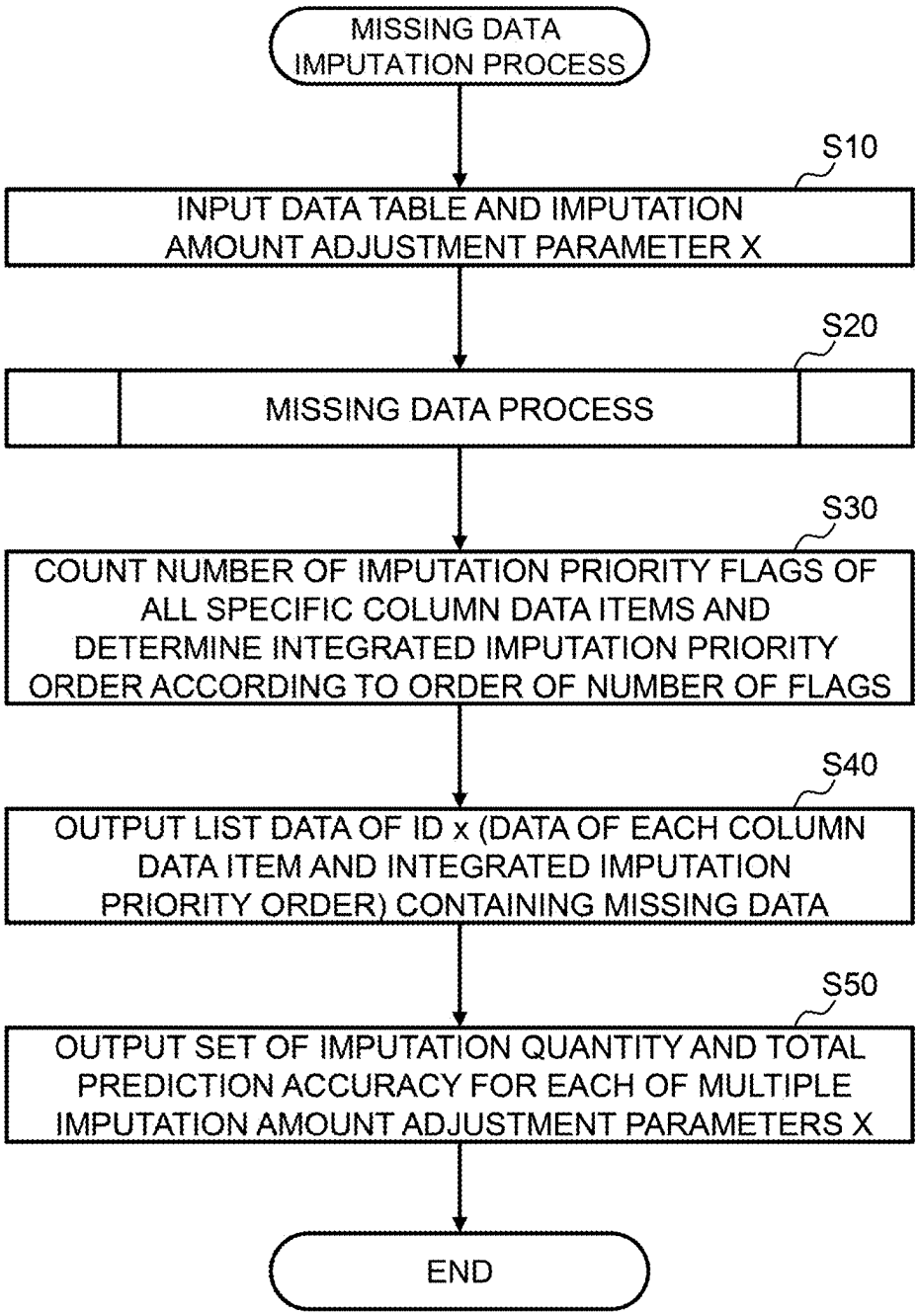

MISSING DATA IMPUTATION PROCESS
S10
INPUT DATA TABLE AND IMPUTATION AMOUNT ADJUSTMENT PARAMETER X
S20
MISSING DATA PROCESS
S30
COUNT NUMBER OF IMPUTATION PRIORITY FLAGS OF ALL SPECIFIC COLUMN DATA ITEMS AND DETERMINE INTEGRATED IMPUTATION PRIORITY ORDER ACCORDING TO ORDER OF NUMBER OF FLAGS
S40
OUTPUT LIST DATA OF ID x (DATA OF EACH COLUMN DATA ITEM AND INTEGRATED IMPUTATION PRIORITY ORDER) CONTAINING MISSING DATA
S50
OUTPUT SET OF IMPUTATION QUANTITY AND TOTAL PREDICTION ACCURACY FOR EACH OF MULTIPLE IMPUTATION AMOUNT ADJUSTMENT PARAMETERS X
END

FIG. 3

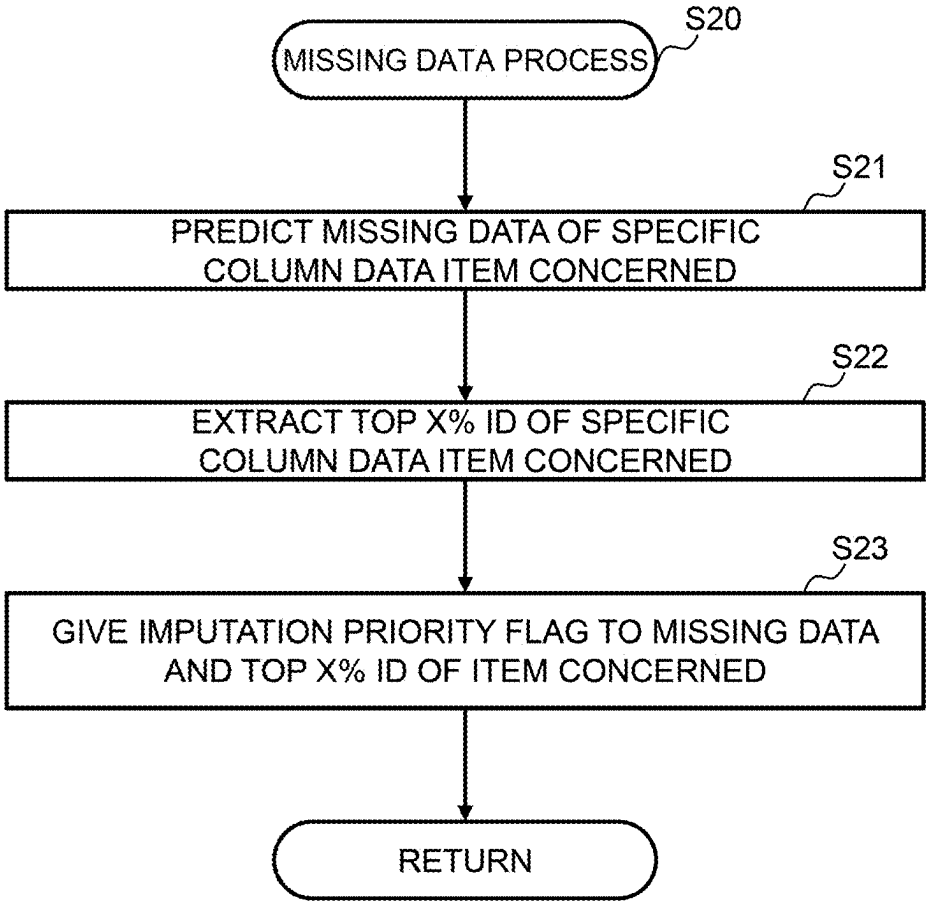
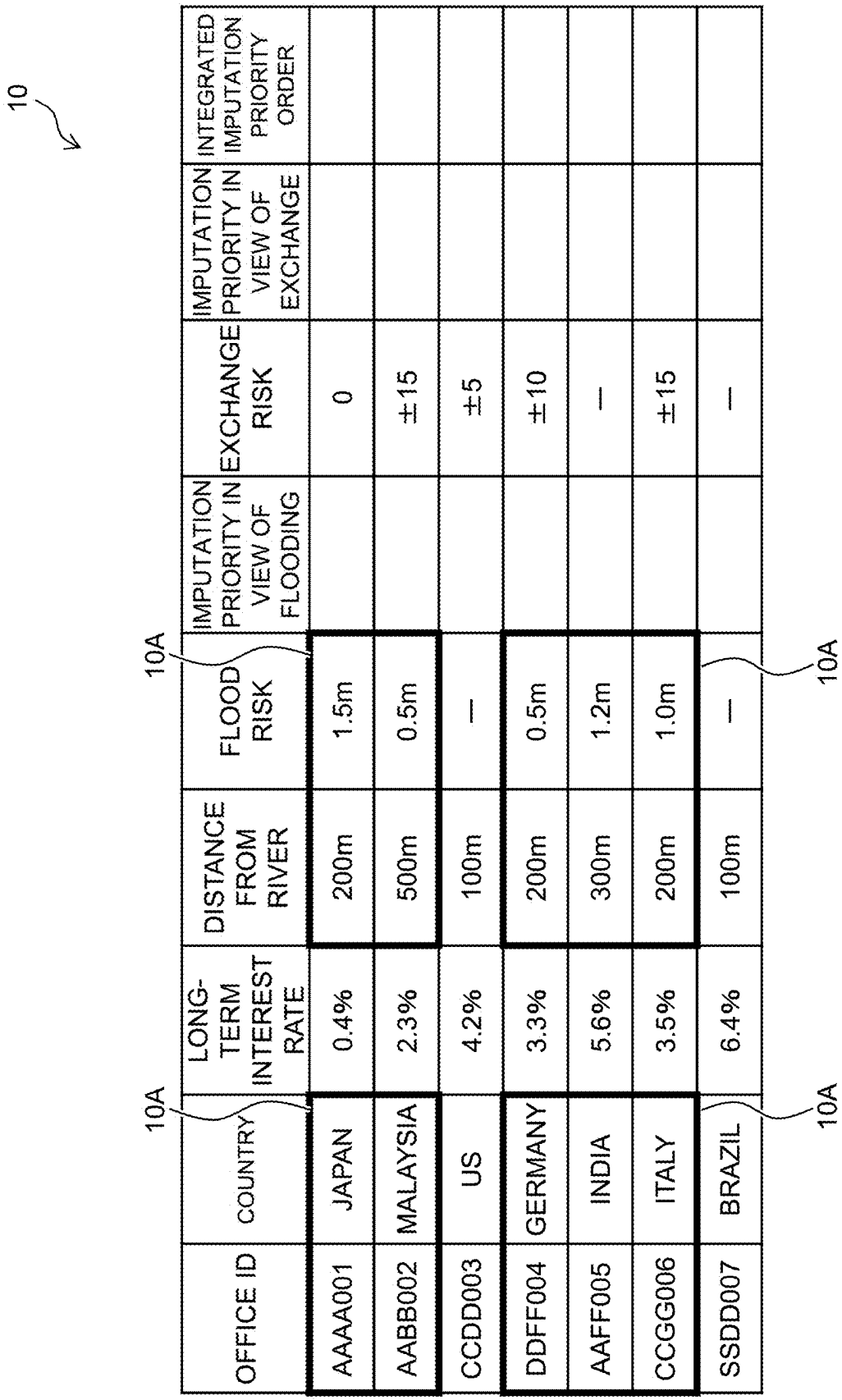

S20
MISSING DATA PROCESS
S21
PREDICT MISSING DATA OF SPECIFIC COLUMN DATA ITEM CONCERNED
S22
EXTRACT TOP X% ID OF SPECIFIC COLUMN DATA ITEM CONCERNED
S23
GIVE IMPUTATION PRIORITY FLAG TO MISSING DATA AND TOP X% ID OF ITEM CONCERNED
RETURN

FIG. 4

10

OFFICE ID	COUNTRY	LONG-TERM INTEREST RATE	DISTANCE FROM RIVER	FLOOD RISK	IMPUTATION PRIORITY IN VIEW OF FLOODING	EXCHANGE RISK	IMPUTATION PRIORITY IN VIEW OF EXCHANGE	INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001	JAPAN	0.4%	200m	1.5m		0		
AABB002	MALAYSIA	2.3%	500m	0.5m		±15		
CCDD003	US	4.2%	100m	—		±5		
DDFF004	GERMANY	3.3%	200m	0.5m		±10		
AAFF005	INDIA	5.6%	300m	1.2m		—		
CCGG006	ITALY	3.5%	200m	1.0m		±15		
SSDD007	BRAZIL	6.4%	100m	—		—		

FIG. 5



OFFICE ID	COUNTRY	LONG-TERM INTEREST RATE	DISTANCE FROM RIVER	FLOOD RISK	IMPUTATION PRIORITY IN VIEW OF FLOODING	EXCHANGE RISK	IMPUTATION PRIORITY IN VIEW OF EXCHANGE	INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001	JAPAN	0.4%	200m	1.5m		0		
AABB002	MALAYSIA	2.3%	500m	0.5m		±15		
CCDD003	US	4.2%	100m	—		±5		
DDFF004	GERMANY	3.3%	200m	0.5m		±10		
AAFF005	INDIA	5.6%	300m	1.2m		—		
CCGG006	ITALY	3.5%	200m	1.0m		±15		
SSDD007	BRAZIL	6.4%	100m	—		—		

FIG. 6

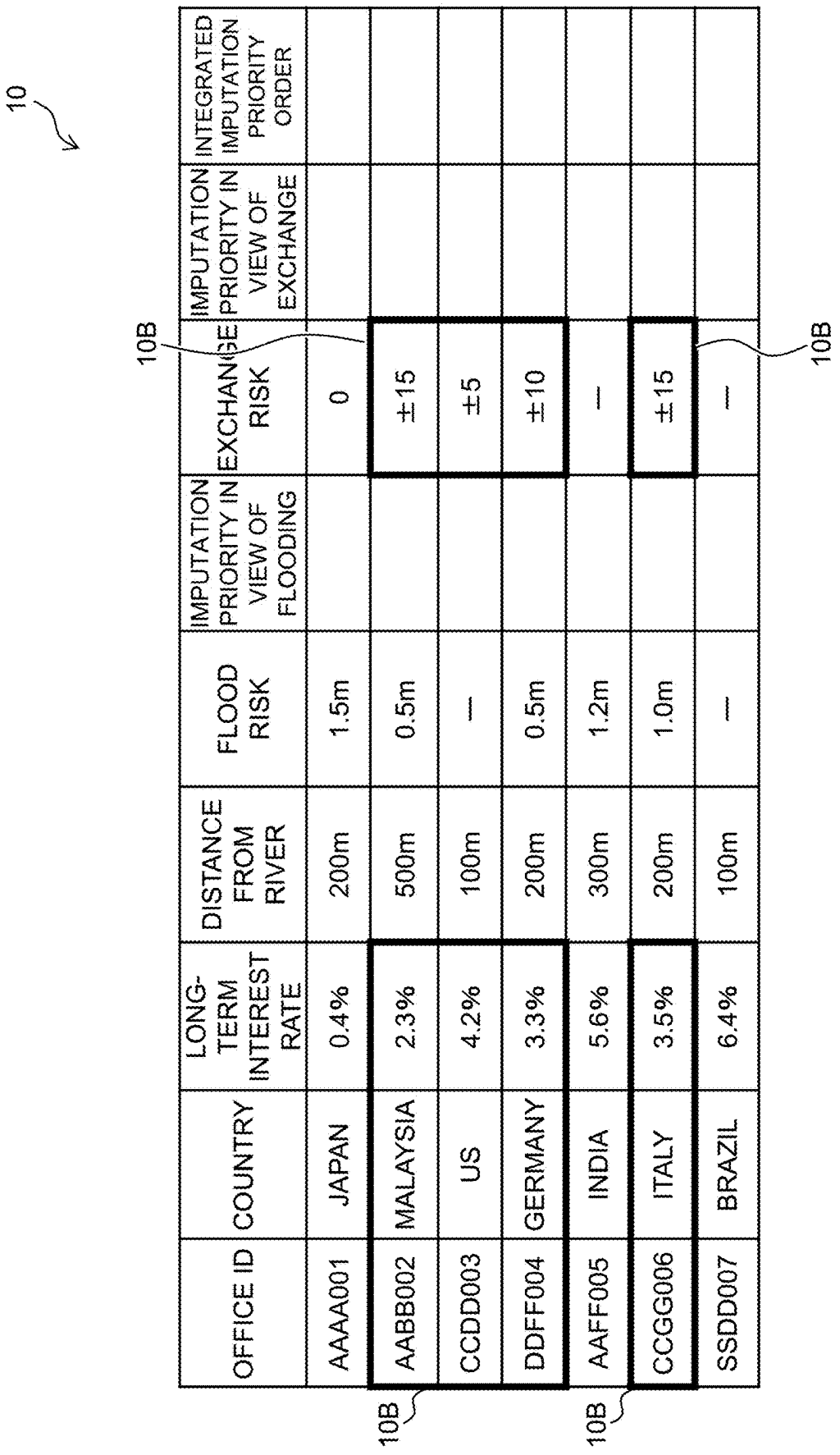

10
10B
10B
10B
OFFICE ID
COUNTRY
LONG-TERM INTEREST RATE
DISTANCE FROM RIVER
FLOOD RISK
IMPUTATION PRIORITY IN VIEW OF FLOODING
EXCHANGE RISK
IMPUTATION PRIORITY IN VIEW OF EXCHANGE
INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001 JAPAN 0.4% 200m 1.5m 0
AABB002 MALAYSIA 2.3% 500m 0.5m ±15
CCDD003 US 4.2% 100m — ±5
DDFF004 GERMANY 3.3% 200m 0.5m ±10
AAFF005 INDIA 5.6% 300m 1.2m —
CCGG006 ITALY 3.5% 200m 1.0m ±15
SSDD007 BRAZIL 6.4% 100m — —

FIG. 7

10

OFFICE ID	COUNTRY	LONG-TERM INTEREST RATE	DISTANCE FROM RIVER	FLOOD RISK	IMPUTATION PRIORITY IN VIEW OF FLOODING	EXCHANGE RISK	IMPUTATION PRIORITY IN VIEW OF EXCHANGE	INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001	JAPAN	0.4%	200m	1.5m		0		
AABB002	MALAYSIA	2.3%	500m	0.5m		±15		
CCDD003	US	4.2%	100m	(PREDICTION) 1.0m		±5		
DDFF004	GERMANY	3.3%	200m	0.5m		±10		
AAFF005	INDIA	5.6%	300m	1.2m		(PREDICTION) ±30		
CCGG006	ITALY	3.5%	200m	1.0m		±15		
SSDD007	BRAZIL	6.4%	100m	(PREDICTION) 1.5m		(PREDICTION) ±20		

FIG. 8

IMPUTATION AMOUNT ADJUSTMENT PARAMETER X = 30%

OFFICE ID	COUNTRY	LONG-TERM INTEREST RATE	DISTANCE FROM RIVER	FLOOD RISK	IMPUTATION PRIORITY IN VIEW OF FLOODING	EXCHANGE RISK	IMPUTATION PRIORITY IN VIEW OF EXCHANGE	INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001	JAPAN	0.4%	200m	1.5m	◎	0	—	
AABB002	MALAYSIA	2.3%	500m	0.5m	—	±15	—	
CCDD003	US	4.2%	100m	(PREDICTION) 1.0m	—	±5	—	
DDFF004	GERMANY	3.3%	200m	0.5m	—	±10	—	
AAFF005	INDIA	5.6%	300m	1.2m	—	(PREDICTION) ±30	◎	
CCGG006	ITALY	3.5%	200m	1.0m	—	±15	—	
SSDD007	BRAZIL	6.4%	100m	(PREDICTION) 1.5m	◎	(PREDICTION) ±20	◎	

10C

ENTER MARK (DOUBLE CIRCLE) IN TOP 30% RISK(⇒**IMPUTATION AMOUNT ADJUSTMENT PARAMETER**) FOR FLOOD RISK→GIVE IMPUTATION PRIORITY (DOUBLE CIRCLE) TO FLOOD RISK OF SSDD007 (BRAZIL-RR)

FIG. 9

10

IMPUTATION AMOUNT ADJUSTMENT PARAMETER X = 30%

OFFICE ID	COUNTRY - DISTRICT	DISTANCE FROM RIVER	FLOOD RISK	IMPUTATION PRIORITY IN VIEW OF FLOODING	EXCHANGE RISK	IMPUTATION PRIORITY IN VIEW OF EXCHANGE	INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001	JAPAN・XX	200m	1.5m	◎	0	—	—
AABB002	MALAYSIA・YY	500m	0.5m	—	±15	—	—
CCDD003	US・QQ	100m	(PREDICTION) 1.0m	—	±5	—	—
DDFF004	GERMANY・SS	200m	0.5m	—	±10	—	—
AAFF005	INDIA・ZZ	300m	1.2m	—	(PREDICTION) ±30	◎	SECOND PLACE
CCGG006	ITALY・WW	200m	1.0m	—	±15	—	—
SSDD007	BRAZIL・RR	100m	(PREDICTION) 1.5m	◎	(PREDICTION) ±20	◎	FIRST PLACE

10D 10D 10D

PROCESS OTHER RISK IN SIMILAR MANNER, AND IMPUTE INFORMATION SEQUENTIALLY FROM ID GIVEN MORE IMPUTATION PRIORITY **FLAGS (RECTANGULAR FRAME)**

FIG. 10

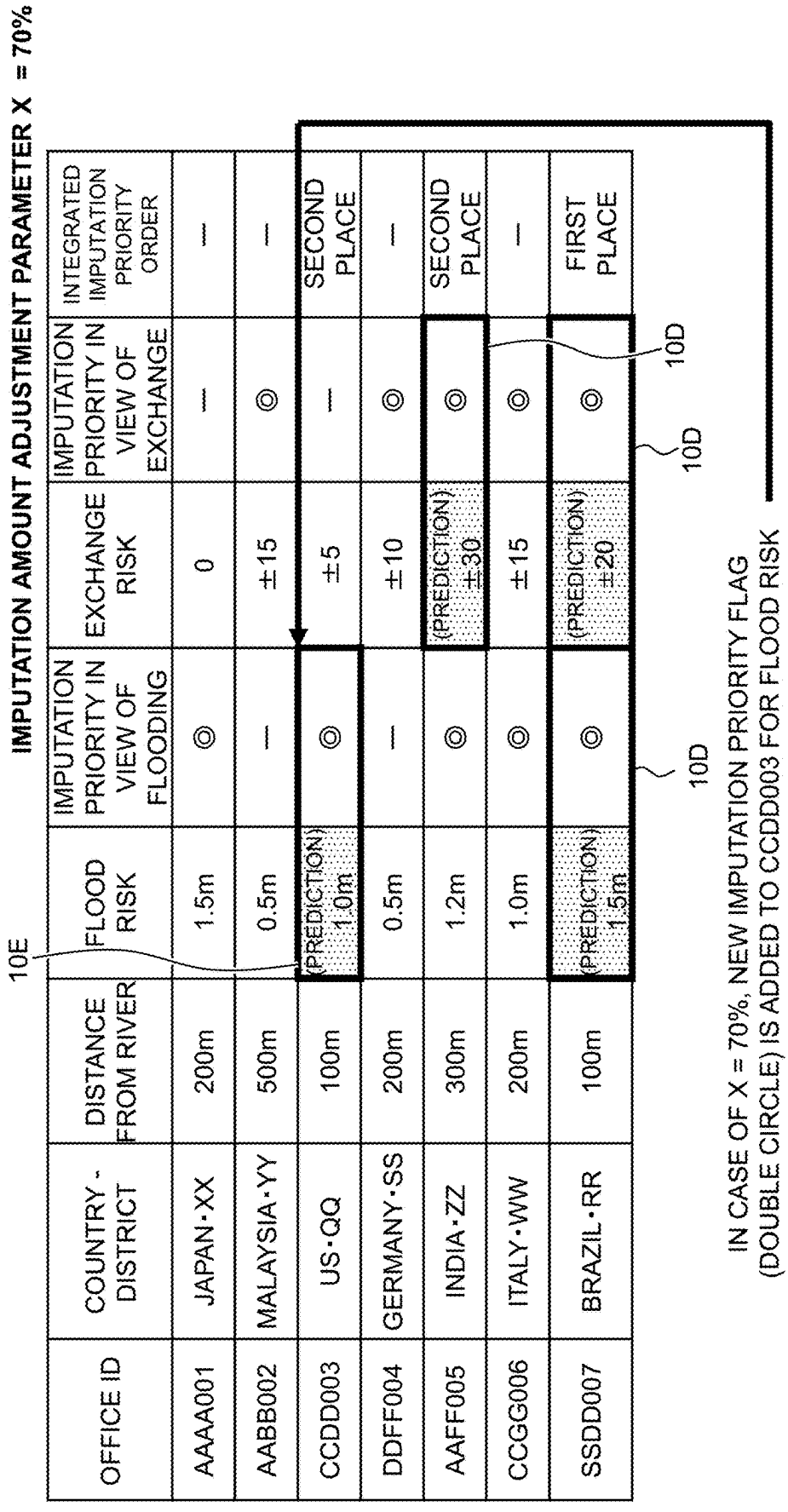
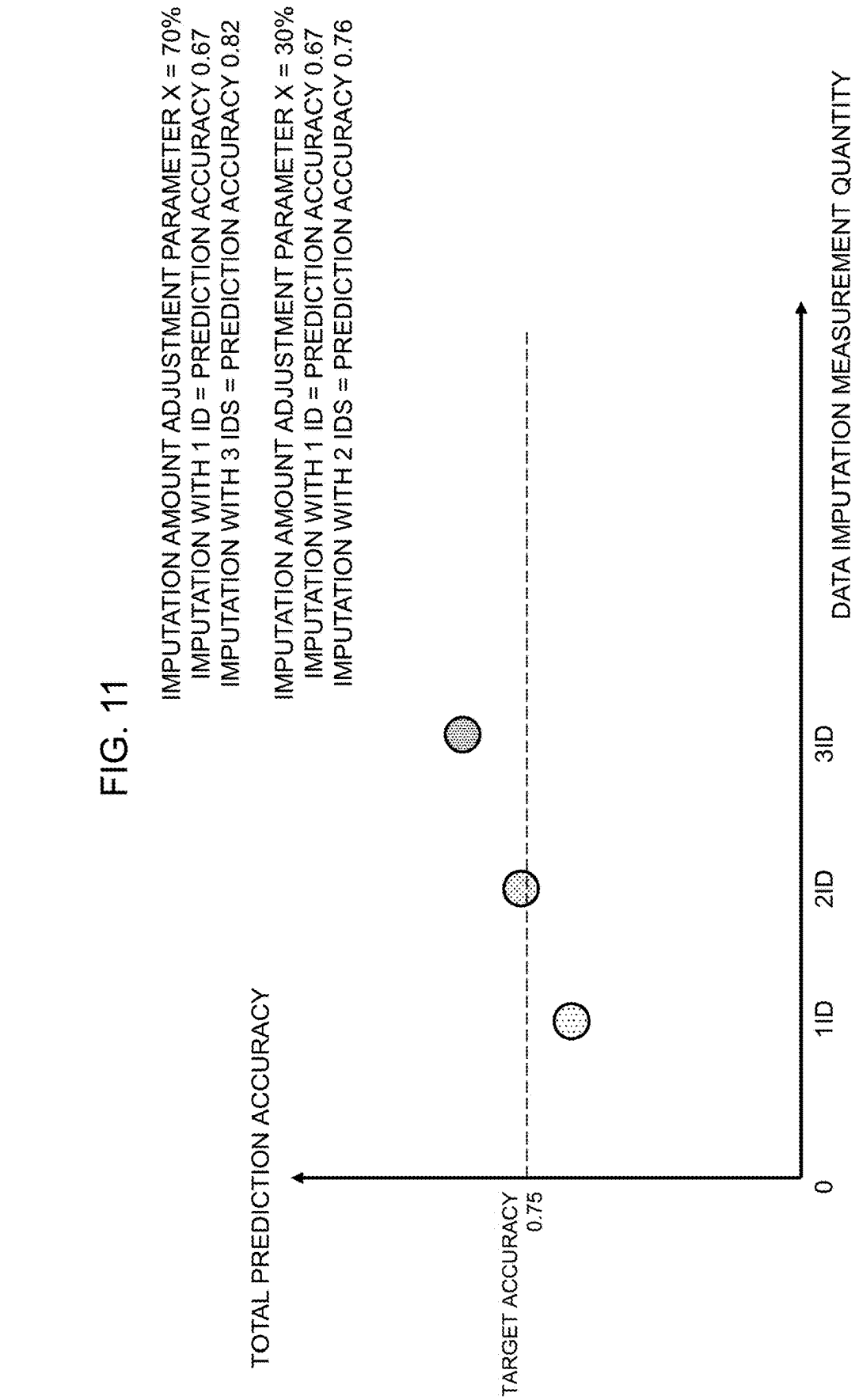

10
10E
IMPUTATION AMOUNT ADJUSTMENT PARAMETER X = 70%
OFFICE ID
COUNTRY-DISTRICT
DISTANCE FROM RIVER
FLOOD RISK
IMPUTATION PRIORITY IN VIEW OF FLOODING
EXCHANGE RISK
IMPUTATION PRIORITY IN VIEW OF EXCHANGE
INTEGRATED IMPUTATION PRIORITY ORDER
AAAA001 JAPAN-XX 200m 1.5m ◎ 0 — —
AABB002 MALAYSIA-YY 500m 0.5m — ±15 ◎ —
CCDD003 US-QQ 100m (PREDICTION) 1.0m ◎ ±5 — SECOND PLACE
DDFF004 GERMANY-SS 200m 0.5m — ±10 ◎ —
AAFF005 INDIA-ZZ 300m 1.2m ◎ (PREDICTION) ±30 ◎ SECOND PLACE
CCGG006 ITALY-WW 200m 1.0m ◎ ±15 ◎ —
SSDD007 BRAZIL-RR 100m (PREDICTION) 1.5m ◎ (PREDICTION) ±20 ◎ FIRST PLACE
10D
10D
10D
IN CASE OF X = 70%, NEW IMPUTATION PRIORITY FLAG (DOUBLE CIRCLE) IS ADDED TO CCDD003 FOR FLOOD RISK PRIORITY

FIG. 11


IMPUTATION AMOUNT ADJUSTMENT PARAMETER X = 70%
IMPUTATION WITH 1 ID = PREDICTION ACCURACY 0.67
IMPUTATION WITH 3 IDS = PREDICTION ACCURACY 0.82
IMPUTATION AMOUNT ADJUSTMENT PARAMETER X = 30%
IMPUTATION WITH 1 ID = PREDICTION ACCURACY 0.67
IMPUTATION WITH 2 IDS = PREDICTION ACCURACY 0.76
TOTAL PREDICTION ACCURACY
TARGET ACCURACY
0.75
0
1ID
2ID
3ID
DATA IMPUTATION MEASUREMENT QUANTITY

MISSING DATA IMPUTATION DEVICE AND MISSING DATA IMPUTATION METHOD

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority based on Japanese patent application No. 2024-022296, filed on Feb. 16, 2024, the entire contents of which are incorporated herein by reference.

BACKGROUND OF THE INVENTION

1. Field of the Invention

[0001] The present invention relates to a missing data imputation device and a missing data imputation method. For example, the present invention is applicable to a missing data imputation device associated with a technology for imputing missing data as a preferable application example.

2. Description of the Related Art

[0002] With recent development of technologies adopting artificial intelligence, studies have been actively conducted concerning machine learning which uses learning data. A part of such learning data contains missing data, in some situations. Accordingly, various studies associated with imputation of missing data have been carried out to deal with these situations. JP-2020-154828-A discloses a technology which imputes missing data as preprocessing of machine learning. The technology disclosed in JP-2020-154828-A is a technology developed to improve imputation accuracy of missing data. Specifically, a correlation matrix calculation unit initially calculates a correlation matrix between attributes by using all learning records. In a case where a missing attribute is present which has a larger absolute value of a correlation value than a correlation threshold, a regression imputation unit performs regression imputation by using the attribute which has the larger absolute value of the correlation value than the correlation threshold.

[0003] Patent Literature: JP-2020-154828-A

[0004] However, as described above, the technology disclosed in JP-2020-154828-A carries out only regression imputation of missing data, and gives no consideration to the extent to which imputation of missing data is required so as to raise total prediction accuracy of learning data containing the missing data.

[0005] The present invention has been developed in consideration of the circumstances described above, and proposes a missing data imputation device and a missing data imputation method each capable of raising total prediction accuracy of learning data containing missing data while more reducing missing data requiring imputation.

SUMMARY OF THE INVENTION

[0006] For solving the problems described above, the present invention includes a data table that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items, and that contains missing data of each of specific column data items in some of the entries, a parameter definition unit that defines an imputation amount adjustment parameter used for adjustment of a proportion requiring imputation in pieces of data of the specific column data items, a missing data prediction unit that predicts the missing data of each of the specific column data items on the basis of data of each of the column data items other than the specific column data items and on the basis of data of each of the specific column data items in the plurality of entries constituting the data table, an extraction unit that extracts at least one of the entries falling within a predetermined proportion defined by the imputation amount adjustment parameter in the respective pieces of data of the specific column data items, an imputation priority flagging unit that gives an imputation priority flag to the at least one entry falling within the predetermined proportion, and a priority order determination unit that counts the number of the entries each given the imputation priority flag in the entries included in the data table, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags.

[0007] In addition, the present invention includes a parameter definition step that causes a parameter definition unit to define an imputation amount adjustment parameter used for adjustment of a proportion requiring imputation in pieces of data of specific column data items in a data table that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items, and that contains missing data of each of the specific column data items in some of the entries, a missing data prediction step that causes a missing data prediction unit to predict the missing data of each of the specific column data items on the basis of data of each of the column data items other than the specific column data items and on the basis of data of each of the specific column data items in the plurality of entries constituting the data table, an extraction step that causes an extraction unit to extract at least one of the entries falling within a predetermined proportion defined by the imputation amount adjustment parameter in the respective pieces of data of the specific column data items, an imputation priority flagging step that causes an imputation priority flagging unit to give an imputation priority flag to the at least one entry falling within the predetermined proportion, and a priority order determination step that causes a priority order determination unit to count the number of the entries each given the imputation priority flag in the entries included in the data table, and to determine an integrated imputation priority order in a descending order of the number of the imputation priority flags.

[0008] According to the present invention, total prediction accuracy of learning data containing missing data can be raised while missing data requiring imputation is further reduced.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] FIG. **1** is a system configuration diagram illustrating a configuration example of a missing data imputation device according to a first embodiment;

[0010] FIG. **2** is a flowchart illustrating an example of procedures for a missing data imputation process;

[0011] FIG. **3** is a flowchart illustrating an example of procedures for a missing data process illustrated in FIG. **2**;

[0012] FIG. **4** is a diagram illustrating an example of details of a data table;

[0013] FIG. **5** is a diagram illustrating an example of the details of the data table;

[0014] FIG. **6** is a diagram illustrating an example of the details of the data table;

[0015] FIG. **7** is a diagram illustrating an example of the details of the data table;

[0016] FIG. **8** is a diagram illustrating an example of the details of the data table;

[0017] FIG. **9** is a diagram illustrating an example of the details of the data table;

[0018] FIG. **10** is a diagram illustrating an example of the details of the data table; and

[0019] FIG. **11** is a diagram illustrating an example of characteristics of total prediction accuracy relative to a data imputation measurement quantity.

DESCRIPTION OF THE PREFERRED EMBODIMENT

[0020] An embodiment according to the present invention will hereinafter be described in detail with reference to the drawings.

[0021] An outline of the present embodiment will be initially described upon. In recent years, creation of highly accurate risk prediction models which use various types of data has been promoted. However, at the time of input of a combination of these pieces of data to the risk prediction models, missing data present in the data causes problems, in some situations. Information available for complementing such missing data is individually collected for some of the models, but such information collection requires a huge workload. Meanwhile, for achieving target prediction accuracy of the risk prediction models, collection of information for complementing all pieces of missing data is not necessarily required. For example, if entries indicated by identifiers (IDs) (hereinafter each simply referred to as "entries") constituting 20 percent of the whole are associated with 80 percent of risks, targeted predictivity can sufficiently be achieved by imputing only missing data of each of these entries constituting 20 percent. Accordingly, the present embodiment gives an imputation priority flag to each of the entries estimated to be associated with 80 percent of the risks, and provides an "integrated imputation priority flag" which indicates an order determined according to the number of the entries each given the imputation priority flag, i.e., according to the number of the imputation priority flags. On the basis of these flags, the present embodiment realizes missing data complementation contributing to high prediction accuracy with a smaller workload. The foregoing points will be hereinafter specifically described.

[0022] FIG. **1** is a system configuration diagram illustrating a configuration example of a missing data imputation device **100** according to a first embodiment. The missing data imputation device **100** is a computer, for example, and includes a data table **10**, a parameter definition unit **20**, a missing data prediction unit **30**, an extraction unit **40**, an imputation priority flagging unit **50**, and a priority order determination unit **60**. The missing data imputation device **100** preferably includes a display unit **70** as well. Note that the display unit **70** may be formed either integrally with or separately from the missing data imputation device **100**.

[0023] The data table **10** includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items. The data table **10** contains missing data of each of specific column data items (flood risk and exchange risk) in some of the entries.

[0024] The parameter definition unit **20** defines an imputation amount adjustment parameter X. The imputation amount adjustment parameter X herein is a parameter used for adjusting a proportion requiring imputation in pieces of data of specific column data items (e.g., a "flood risk" and an "exchange risk" described below). According to the present embodiment, a predetermined proportion, such as top 30%, or preferably top 70%, is defined as the imputation amount adjustment parameter X.

[0025] The missing data prediction unit **30** predicts missing data of each of the specific column data items on the basis of data of each of the column data items other than the specific column data items and on the basis of data of each of the specific column data items in the plurality of entries constituting the data table **10**.

[0026] For example, the missing data prediction unit **30** learns a model on the basis of an explanatory variable as the data of each of the column data items other than the specific column data items and on the basis of an objective variable as the data of each of the specific column data items in the plurality of entries constituting the data table **10**. The missing data prediction unit **30** predicts the missing data of each of the specific column data items through machine learning by using this model.

[0027] The extraction unit **40** extracts at least one entry falling within the predetermined proportion defined by the imputation amount adjustment parameter X in the respective pieces of data of the specific column data items.

[0028] The imputation priority flagging unit **50** gives an imputation priority flag to the extracted entry described above, i.e., at least one entry falling within the predetermined proportion defined by the imputation amount adjustment parameter X in the data table **10**.

[0029] The priority order determination unit **60** counts the number of the entries each given the imputation priority flag in the entries included in the data table **10**, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags given to the entries.

[0030] The display unit **70** displays list data containing data of each of the specific column data items and the integrated imputation priority order for each of the entries containing missing data. Details of this display will be described below.

[0031] The imputation priority flagging unit **50** provides the imputation priority flag for each of a plurality of the foregoing predetermined proportions defined as the imputation amount adjustment parameters (e.g., 30% and 70%). The display unit **70** displays a data imputation measurement quantity based on the list data described above and prediction accuracy associated with the data of each of the specific column data items (e.g., the "flood risk" and the "exchange risk" described below) and obtained by the missing data prediction unit **30**.

[0032] Subsequently described will be a missing data imputation method as an operation example of the missing data imputation device **100** configured as above. This missing data imputation method includes: a parameter definition step that causes the parameter definition unit **20** to define the imputation amount adjustment parameter X used for adjustment of a proportion requiring imputation in pieces of data

of specific column data items in the data table **10** that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items and that contains missing data of each of the specific column data items in some of the entries; a missing data prediction step that causes the missing data prediction unit **30** to predict the missing data of each of the specific column data items on the basis of data of each of the column data items other than the specific column data items and on the basis of data of each of the specific column data items (e.g., pieces of data excluding missing data) in the plurality of entries constituting the data table **10**; an extraction step that causes the extraction unit **40** to extract at least one of the entries falling within a predetermined proportion defined by the imputation amount adjustment parameter X in the respective pieces of data of the specific column data items; an imputation priority flagging step that causes the imputation priority flagging unit **50** to give an imputation priority flag to the at least one entry falling within the predetermined proportion; and a priority order determination step that causes the priority order determination unit **60** to count the number of the entries each given the imputation priority flag in the entries included in the data table **10** and to determine an integrated imputation priority order in a descending order of the number of the imputation priority flags.

[0033] FIG. **2** is a flowchart illustrating an example of procedures for the missing data imputation process, and FIG. **3** is a flowchart illustrating an example of procedures for a missing data process illustrated in FIG. **2** (step S**20** in FIG. **2**). Description will be hereinafter presented with reference to the data table **10** illustrated in FIGS. **4** to **9**.

[0034] In step S**10** in FIG. **2**, the data table **10** and the imputation amount adjustment parameter X are input to the missing data imputation device **100**. The parameter definition unit **20** sets a proportion of processing to top 30%, for example, on the basis of the input imputation amount adjustment parameter X.

[0035] Meanwhile, for example, the data table **10** defines column data items such as an office ID, a country, a long-term interest rate, a distance from a river, a flood risk, an imputation priority in view of flooding, an exchange risk, an imputation priority in view of exchange, and an integrated imputation priority order, as illustrated in FIG. **4**. The data table **10** includes respective pieces of data for these column data items, and manages entries identifiable from each other on the basis of the office ID described above. For example, the data table **10** illustrated in FIG. **4** contains missing data expressed as "-" for the flood risk and the exchange risk as examples of the specific column data items.

[0036] In step S**20**, the missing data process is executed. This missing data process is carried out by the missing data prediction unit **30**. In step S**21** in FIG. **3**, the missing data prediction unit **30** predicts data of each of these specific column data items (a missing value expressed as "–" in the figure).

[0037] Specifically, in the case of the flood risk, for example, the missing data prediction unit **30** creates a model for predicting the specific column data item containing missing data, such as data of the flood risk, on the basis of a combination of the entries containing no missing data. For example, the combination of the entries containing no missing data herein refers to the entries of the office IDs "AAAA001," "AABB002," "DDFF004," "AAFF005," and "CCGG006," indicated by bold lines **10**A in FIG. **5**. It is assumed in the present embodiment that the flood risk=S is influenced by the country and the distance from the river. Note that these two items (the country and the distance from the river) are explanatory variables and that the flood risk is an objective variable in the present embodiment.

[0038] Meanwhile, in the case of the exchange risk, for example, the missing data prediction unit **30** creates a model for predicting the specific column data item containing missing data, for example, data of the exchange risk, on the basis of a combination of the entries containing no missing data. For example, the combination of the entries containing no missing data herein refers to the entries of the office IDs "AAAA002," "CCDD003," "DDFF004," and "CCGG006," indicated by bold lines **10**B in FIG. **6**. It is assumed in the present embodiment that the exchange risk=K is influenced by the long-term interest rate of the country. Note that this item (the long-term interest rate of the country) is an explanatory variable and that the exchange risk is an objective variable in the present embodiment.

[0039] For example, the missing data prediction unit **30** learns a model on the basis of an explanatory variable as the data of each of the column data items other than the specific column data items and on the basis of an objective variable as the data of each of the specific column data items (e.g., flood risk and exchange risk) in the plurality of entries constituting the data table **10**, and predicts missing data of each of the specific column data items through machine learning by using this model (corresponding to parts of "prediction" in FIG. **7**).

[0040] In step S**22** in FIG. **3**, the extraction unit **40** extracts at least one entry falling within the predetermined proportion (e.g., top 30%) defined by the imputation amount adjustment parameter X in the respective pieces of data of the specific column data items (e.g., flood risk and exchange risk).

[0041] In step S**23** in FIG. **3**, the imputation priority flagging unit **50** gives an imputation priority flag to at least one entry falling within the predetermined proportion (e.g., top 30%) defined by the imputation amount adjustment parameter X (see FIG. **8**). Note that a double circle is entered in each entry to which the imputation priority flag has been given in the example illustrated in the figure.

[0042] In subsequent step S**30** in FIG. **2**, the priority order determination unit **60** counts the number of the entries each given the imputation priority flag in the entries included in the data table **10**, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags (see FIG. **9**).

[0043] In step S**40**, the display unit **70** displays list data of the entry x containing missing data (data of each column data item and integrated imputation priority order).

[0044] In step S**50**, the display unit **70** displays a set of an imputation quantity and total prediction accuracy for each of a plurality of the imputation amount adjustment parameters X.

[0045] FIG. **10** is a diagram illustrating an example of the missing data imputation process performed under the imputation amount adjustment parameter X of 70%. In the description of the example illustrated in FIG. **10**, points similar to the corresponding points described above in the example illustrated in FIGS. **8** and **9** will not be repeatedly explained.

[0046] The imputation amount adjustment parameter X in the example illustrated in FIG. **10** is set to 70%. The extraction unit **40** extracts at least one entry falling within a range of top 70% in respective pieces of data of specific column data items (e.g., flood risk and exchange risk).

[0047] The imputation priority flagging unit **50** gives an imputation priority flag not only to data **10**D extracted in FIG. **9** described above, but also to data **10**E of the flood risk associated with the entry of the office ID "CCDD003" (corresponding to the double circle for the imputation priority of the flood risk in the figure).

[0048] The priority order determination unit **60** counts the number of the entries each given the imputation priority flag in the entries included in the data table **10**, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags. As a result, the integrated imputation priority order is determined as follows. The entry of the office ID "SSDD007" having two double circles holds the first place, the entry of the office ID "CCDD003" having one double circle holds the second place, and the entry of the office ID "AAFF005" having one double circle holds the third place. Accordingly, the example illustrated in FIG. **10** is different from the example illustrated in FIG. **9** in that the entry of the office ID "CCDD003" is also covered.

[0049] FIG. **11** is a diagram illustrating an example of characteristics of total prediction accuracy relative to a data imputation measurement quantity. The example illustrated in the figure indicates a case in which the imputation amount adjustment parameter X is set to 70% and a case in which the imputation amount adjustment parameter X is set to 30%, in a state where targeted prediction accuracy (corresponding to "target accuracy" in the figure) is 0.75. According to the example illustrated in the figure, "1 ID" in the figure indicates imputation using an entry corresponding to one office ID (hereinafter also referred to as "1 ID imputation"), "2 IDs" indicates imputation using entries corresponding to two office IDs (hereinafter also referred to as "2 ID imputation"), and "3 IDs" indicates imputation using entries corresponding to three office IDs (hereinafter also referred to as "3 ID imputation").

[0050] It is intended in this example that the number of entries corresponding to the office ID requiring imputation is determined in consideration of the target accuracy described above. According to the present embodiment, the missing data prediction unit **30** examines which value is to be set as the value of the imputation amount adjustment parameter X (30% or 70% herein) so as to obtain prediction accuracy of "0.75" as one example of target accuracy. In other words, the missing data prediction unit **30** selects, from a plurality of predetermined proportions each defined as the imputation amount adjustment parameter X (e.g., 30% and 70%), one predetermined proportion appropriate for obtaining sufficient total prediction accuracy of learning data containing missing data for meeting the target accuracy.

[0051] The example illustrated in the figure is characterized in that the data imputation measurement quantity increases in the order of "1 ID," "2 IDs," and "3 IDs" and that the total prediction accuracy improves according to the increase in the data imputation measurement quantity. For example, the prediction accuracy of "2 IDs" is 0.76, the prediction accuracy of "1 ID" is lower than 0.75 (e.g., 0.67), and the prediction accuracy of "3 IDs" exceeds 0.75 (e.g., 0.82).

[0052] According to the present embodiment, for obtaining prediction accuracy reliably reaching the target accuracy of "0.75" described above under the imputation amount adjustment parameter X of 30%, 2 ID imputation is selected instead of 1 ID imputation which does not provide sufficient prediction accuracy.

[0053] Meanwhile, for obtaining prediction accuracy reliably reaching the target accuracy of "0.75" described above under the imputation amount adjustment parameter X of 70% in the present embodiment, 3 ID imputation is selected instead of 1 ID imputation which does not provide sufficient prediction accuracy.

[0054] As described above, the missing data prediction unit **30** sets the imputation amount parameter X to a value sufficient for obtaining prediction accuracy corresponding to the target accuracy "0.75" described above. In this manner, the missing data prediction unit **30** can raise total prediction accuracy of learning data containing missing data by using the model described above.

[0055] As described above, the missing data imputation device **100** according to the present embodiment includes: the data table **10** that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items and that contains missing data of each of specific column data items (e.g., flood risk and exchange risk) in some of the entries; the parameter definition unit **20** that defines the imputation amount adjustment parameter X used for adjustment of a proportion requiring imputation in pieces of data of the specific column data items; the missing data prediction unit **30** that predicts the missing data of each of the specific column data items on the basis of data of each of the column data items other than the specific column data items and on the basis of data of each of the specific column data items in the plurality of entries constituting the data table **10**; the extraction unit **40** that extracts at least one of the entries falling within a predetermined proportion (e.g., 30%) defined by the imputation amount adjustment parameter X in the respective pieces of data of the specific column data items; the imputation priority flagging unit **50** that gives an imputation priority flag to the at least one entry falling within the predetermined proportion; and the priority order determination unit **60** that counts the number of the entries each given the imputation priority flag in the entries included in the data table **10**, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags.

[0056] As described above, at least one entry falling within the predetermined proportion defined by the imputation amount adjustment parameter X is extracted by the extraction unit **40**, and the imputation priority flag is given to the extracted entry to determine the integrated imputation priority order in a descending order of the number of the imputation priority flags. In this case, missing data deeply influencing prediction accuracy is precisely recognizable with reference to the integrated imputation priority order. Accordingly, total prediction accuracy of learning data containing missing data can be raised while missing data requiring imputation is further reduced.

[0057] The missing data imputation device **100** according to the present embodiment includes the display unit **70** which displays list data containing the data of each of the specific column data items and the integrated imputation

priority order for each of the entries containing missing data. With reference to the list data containing the data of each of the specific column data items and the integrated imputation priority order for each of the entries containing missing data in this configuration, total prediction accuracy of learning data containing missing data can be raised while missing data requiring imputation is further reduced.

[0058] According to the present embodiment, the imputation priority flagging unit **50** provides imputation priority flags for each of a plurality of predetermined proportions (e.g., 30% and 70%) defined as the imputation amount parameters X, while the display unit **70** displays a data imputation measurement quantity based on the list data, and prediction accuracy associated with data of each of the specific column data items and obtained by the missing data prediction unit **30**. With reference to the data imputation measurement quantity based on the list data and the prediction accuracy associated with the data of each of the specific column data items and obtained by the missing data prediction unit **30** in this configuration, total prediction accuracy of learning data containing missing data can be raised while missing data requiring imputation is further reduced.

[0059] According to the present embodiment, the missing data prediction unit **30** selects, from a plurality of predetermined proportions each defined as the imputation amount adjustment parameter X, one predetermined proportion appropriate for obtaining sufficient total prediction accuracy of learning data containing missing data for meeting the target accuracy. In this manner, total prediction accuracy of learning data containing missing data can be raised to target accuracy or higher while missing data requiring imputation is further reduced.

[0060] According to the present embodiment, the missing data prediction unit **30** learns a model on the basis of an explanatory variable as data of each of the column data items other than specific column data items in a plurality of entries constituting the data table **10** and on the basis of an objective variable as data of each of the specific column data items (e.g., the flood risk and the exchange risk described above), and predicts missing data of each of the specific column data items through machine learning by using this model. In this manner, the missing data prediction unit **30** can predict missing data by using the model, in a state in which total prediction accuracy of learning data containing missing data is enhanced while further reduction of missing data requiring imputation is achieved.

[0061] Note that the present invention is not limited to the embodiment described above and includes various modifications and equivalent configurations within the scope of the appended claims. For example, the above embodiment has been explained in detail for easy understanding of the present invention. Accordingly, the present invention is not necessarily required to be equipped with all the configurations described above. In addition, elements presented in parallel in the description of the present embodiment may be configured such that at least one of the elements is connected to another element or the other elements in series.

[0062] The present invention is applicable to a missing data imputation device associated with a technology for imputing missing data.

What is claimed is:

1. A missing data imputation device comprising:

a data table that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items, and that contains missing data of each of specific column data items in some of the entries;

a parameter definition unit that defines an imputation amount adjustment parameter used for adjustment of a proportion requiring imputation in pieces of data of the specific column data items;

a missing data prediction unit that predicts the missing data of each of the specific column data items on a basis of data of each of the column data items other than the specific column data items and on a basis of data of each of the specific column data items in the plurality of entries constituting the data table;

an extraction unit that extracts at least one of the entries falling within a predetermined proportion defined by the imputation amount adjustment parameter in the respective pieces of data of the specific column data items;

an imputation priority flagging unit that gives an imputation priority flag to the at least one entry falling within the predetermined proportion; and

a priority order determination unit that counts the number of the entries each given the imputation priority flag in the entries included in the data table, and determines an integrated imputation priority order in a descending order of the number of the imputation priority flags.

2. The missing data imputation device according to claim **1**, further comprising:

a display unit that displays list data containing the data of each of the specific column data items and the integrated imputation priority order for each of the entries containing the missing data.

3. The missing data imputation device according to claim **2**, wherein

the imputation priority flagging unit gives the imputation priority flag to each of a plurality of the predetermined proportions defined as the imputation amount adjustment parameters, and

the display unit displays a data imputation measurement quantity based on the list data, and prediction accuracy associated with the data of each of the specific column data items and obtained by the missing data prediction unit.

4. The missing data imputation device according to claim **3**, wherein

the missing data prediction unit selects, from a plurality of the predetermined proportions defined as the imputation amount adjustment parameters, one predetermined proportion appropriate for obtaining sufficient total prediction accuracy of learning data containing the missing data for meeting target accuracy.

5. The missing data imputation device according to claim **1**, wherein

the missing data prediction unit learns a model on a basis of an explanatory variable as the data of each of the column data items other than the specific column data items and on a basis of an objective variable as the data of each of the specific column data items in the plurality of entries constituting the data table, and predicts the missing data of each of the specific column data items through machine learning by using the model.

6. A missing data imputation method comprising:

a parameter definition step that causes a parameter definition unit to define an imputation amount adjustment parameter used for adjustment of a proportion requiring imputation in pieces of data of specific column data items in a data table that includes a plurality of column data items defined in a column direction and a plurality of entries defined in a row direction and each including respective pieces of data of the plurality of column data items, and that contains missing data of each of the specific column data items in some of the entries;

a missing data prediction step that causes a missing data prediction unit to predict the missing data of each of the specific column data items on a basis of data of each of the column data items other than the specific column data items and on a basis of data of each of the specific column data items in the plurality of entries constituting the data table;

an extraction step that causes an extraction unit to extract at least one of the entries falling within a predetermined proportion defined by the imputation amount adjustment parameter in the respective pieces of data of the specific column data items;

an imputation priority flagging step that causes an imputation priority flagging unit to give an imputation priority flag to the at least one entry falling within the predetermined proportion; and

a priority order determination step that causes a priority order determination unit to count the number of the entries each given the imputation priority flag in the entries included in the data table, and to determine an integrated imputation priority order in a descending order of the number of the imputation priority flags.

7. The missing data imputation method according to claim 6, further comprising:

a display step that causes a display unit to display list data containing the data of each of the specific column data items and the integrated imputation priority order for each of the entries containing the missing data.

8. The missing data imputation method according to claim 7, wherein

the imputation priority flagging unit gives the imputation priority flag to each of a plurality of values defined as the imputation amount adjustment parameters, and

the display step causes the display unit to display a data imputation measurement quantity based on the list data, and prediction accuracy associated with the data of each of the specific column data items and obtained by the missing data prediction unit.

9. The missing data imputation method according to claim 8, wherein

the missing data prediction step causes the missing data prediction unit to select, from a plurality of the predetermined proportions defined as the imputation amount adjustment parameters, one predetermined proportion appropriate for obtaining sufficient total prediction accuracy of learning data containing the missing data for meeting target accuracy.

10. The missing data imputation method according to claim 6, wherein

the missing data prediction step causes the missing data prediction unit to learn a model on a basis of an explanatory variable as the data of each of the column data items other than the specific column data items and on a basis of an objective variable as the data of each of the specific column data items in the plurality of entries constituting the data table, and to predict the missing data of each of the specific column data items through machine learning by using the model.

* * * * *