



Linear correlation discovery in databases: a data mining approach

Roger H.L. Chiang ^{a,*}, Chua Eng Huang Cecil ^b, Ee-Peng Lim ^c

^a *Information Systems Department, College of Business, University of Cincinnati, P.O. Box 210211, Cincinnati, OH 45221-0211, USA*

^b *Nanyang Business School, Nanyang Technological University, Singapore 639798, Singapore*

^c *Center for Advanced Information Systems, School of Computer Engineering, Nanyang Technological University, Singapore 639798, Singapore*

Received 8 September 2004; accepted 8 September 2004

Available online 19 October 2004

Abstract

Very little research in knowledge discovery has studied how to incorporate statistical methods to automate linear correlation discovery (LCD). We present an automatic LCD methodology that adopts statistical measurement functions to discover correlations from databases' attributes. Our methodology automatically pairs attribute groups having potential linear correlations, measures the linear correlation of each pair of attribute groups, and confirms the discovered correlation. The methodology is evaluated in two sets of experiments. The results demonstrate the methodology's ability to facilitate linear correlation discovery for databases with a large amount of data.

© 2004 Elsevier B.V. All rights reserved.

Keywords: Knowledge discovery in database; Linear correlation; Association measurement; Data mining

* Corresponding author. Tel.: +1 513 556 7086; fax: +1 513 556 4891.
E-mail address: roger.chiang@uc.edu (R.H.L. Chiang).

1. Introduction

As competition among businesses continue to increase, it is crucial for organizations to discover knowledge that could give them advantages over their competitors. In the past, such knowledge often was obtained by collecting data and testing it against some predefined hypothesis (i.e., a hypothetico-deductive approach to obtaining knowledge). Lately, greater emphasis has been given to discovering (inducing) knowledge from existing databases. The knowledge discovery approach employs various data mining algorithms such as association rule mining algorithms [1] to obtain knowledge from databases. However, very little work has investigated the possibility of automating traditional data analysis using statistical methods for knowledge discovery [2–5]. Because the amount of data generated and accumulated continues to exceed the number of available experienced analysts [6], it is imperative to develop methods to automate and expedite data analysis for knowledge discovery from existing databases [7,8]. This research establishes a novel discovery methodology to induce business knowledge (also called business intelligence) in the form of linear correlations for better decision making.

As an illustration of the usefulness of such automated knowledge discovery, consider an organization with globally distributed factories that wants to determine how factory effectiveness can be improved. Factory effectiveness includes various aspects, such as, cost per unit produced, output per day and factory downtime. Furthermore, many possible factors could influence factory effectiveness, including wages, reliability of supply, and age of the factory. In traditional data analysis, data analysts must first propose a set of hypotheses for testing. Statistics software, such as SAS/STAT and SPSS Base, can provide only the mechanisms to test these possible relationships [9]. Therefore, data analysts are responsible for ascertaining the appropriate analysis that will identify relationships through hypothesis testing, and must manually select the appropriate factor, outcome, and measurement function for each analysis. Often, fatigue, overhead, manpower cost, and the limits of human cognitive capability detract from a thorough understanding and complete analysis of the sheer volume of data available from existing business databases.

In this research, we demonstrate that these manual tasks of traditional data analysis (i.e., proposing hypotheses and selecting factors, outcomes, and measurement functions) can be automated for knowledge discovery in databases. Specifically, we automate linear correlation discovery (LCD), the goal of which is to determine whether two attributes or sets of attributes (i.e., attribute groups) have a relationship. A thorough discussion of LCD is presented in Section 2.1.

Some previous work has addressed related problems. Hou [3] developed a system that determined whether a regression or a classifier were appropriate for analyzing a system. The SNOOT project [2] derived some properties of attributes that could be leveraged for analysis. Aladwani [4] created an expert system to select an appropriate multiple-comparison test. Some authors have developed clustering or classification algorithms based on correlation (e.g., derivatives of Principle Component Analysis [5]), while others have developed pre-processors to select the ‘best’ algorithm for a given task [7,8]. However, to our knowledge, no one has attempted specifically to automate linear correlation discovery.

1.1. Research objective and contributions

The objective of this research is to demonstrate the feasibility of automating and expediting the LCD process. To do so, we develop a methodology that contributes to knowledge discovery in databases, particularly LCD, in four ways:

- *Classify attributes.* We semi-automatically derive the measurement properties of attributes such as distance and order and employ this information to classify attributes. Attribute classification occurs through the analysis of schema information, such as attribute data type and length, and data contents (attribute values) of the target relational database. Although users and analysts may be able to perform this task manually, our goal is to automate as much of the LCD process as possible to reduce unnecessary human involvement.
- *Consider attribute groups.* We consider potential correlations among sets of attributes (i.e., attribute groups) as well as among individual attributes. For example, it is not sufficient to conjecture that the age of a factory or employee wage alone influences factory performance. It is instead necessary to conjecture that factory age and wages together affect factory performance.
- *Determine the correlation measurement functions.* We establish a set of heuristic rules to determine the most appropriate correlation measurement function to measure each potential induced linear correlation.
- *Confirm the discovered correlations.* The discovered linear correlations are confirmed by repeating the same measurement on subsequent data samples. Therefore, artifact discoveries can be rejected automatically.

To evaluate the proposed methodology, we have developed a prototype system, the Linear Correlation Discovery System. The prototype uses the Visual Basic language as well as MS Access and SPSS Base as the underlying database and statistical analysis packages, respectively [10].

1.2. Paper organization

The remainder of this paper is organized as follows. Section 2 introduces LCD and provides an overview of the proposed methodology. Section 3 discusses the use of random samples to expedite the discovery process. Sections 4 and 5 discuss the induction and measurement of potential linear correlations, respectively. Section 6 presents the method to confirm discovered correlations. Section 7 then elaborates on the evaluation experiments. Section 8 concludes the paper and discusses further directions. Appendices A and B present the heuristic rules established for the proposed methodology.

2. Linear correlation discovery methodology

The LCD methodology requires that data for discovery have two characteristics. First, they must be represented in tabular form because most quantitative analyses assume a tabular representation of data [11] and most business databases currently are based on the relational data model (i.e., tabular form). Second, the data must have four common data types: Integer, Decimal,

Table 1
Attributes of example table for linear correlation discovery

| Attribute | Description |
|-------------------|---|
| Factory_ID | The ID number of the factory |
| Region | The continent on which the factory is located |
| Country | The country in which the factory is located |
| Worker_Age | Average age of the factory's workers |
| Avg_Overtime | Average amount of overtime of workers |
| Avg_Wage | Average wage of workers |
| Start_Date | Date the factory began operation |
| Transport_Service | 1 means the transport company supplying the factory is completely dependent on the company for business. 2 means reliable transport resources 3 means unreliable transport resources |
| Has_ERP | Whether the factory is connected to the organizational ERP |
| Cost_Unit | The cost to make one widget in the factory |
| Widgets_Made | Number of widgets made per year |
| Defects | Number of defective widgets per 1000 widgets made |
| Factory_Downtime | Number of hours per year factory was inoperative |

Date, and String. Some data types such as Varchar, Memo, Currency and Boolean can be mapped as String, String, Decimal and Integer respectively. Other data types (e.g., Graphic, OLE_object) that require graphics, document matching, or other sophisticated functions to measure their correlation are not considered.

Throughout this paper, we illustrate the LCD methodology using an example table from a fictitious global corporation that sells only one kind of item: widgets. Table 1 presents the attributes of this example.

2.1. Linear correlation discovery

Linear correlation discovery refers to the discovery of associations among attribute groups (sets of attributes). The term “linear” refers to the assumption that all attributes within an attribute group are independent of one another and their effects are therefore additive. For example, assume a relationship between *Worker_Age* and *Factory_Age* on *Widgets_Made*. If factory B is half as old as factory A and has workers that are half as old as those in factory A, a linear method for data analysis would assume that factory A is half as productive as factory B. The proposed LCD methodology does not simply assume that factory A was one-fourth as productive as factory B (i.e., $\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$).¹

“Correlation” refers to the proportion of explained variance between two attribute groups [12] and attempts to determine whether the dispersion (variance) of values in one attribute group can predict the dispersion in another. For example, the workers in different factories will have separate ages. Similarly, individual factories will have their own levels of performance. If younger (older)

¹ Note that linearity is a conservative assumption. An LCD analysis can identify that a relationship exists between the attributes in a multiplicative (interaction) case but would specify the relationship incorrectly and misestimate the *strength* and *significance* of the relationship.

workers tend to improve (reduce) factory productivity, there is a correlation between worker age and factory productivity.

Linear correlation discovery differs from association rule mining or market basket analysis [1,13,14], in that it identifies relationships among attribute groups and is suited to making general inferences about a domain. Association rule discovery, in contrast, identifies the relationships among the attributes' instances and is useful for deriving local properties of the instances. For example, the widely accepted idea that workers who have more than eight hours overtime per week tend to have higher defect rates is an instance-based relationship and can be supported with association rule mining techniques. Conversely, the general rule that overtime is inversely related to productivity is discoverable only through a correlation. Therefore, LCD is more useful than association rule mining for inferring global tendencies about attribute groups; it is, for example, more useful to develop a general rule that relates overtime and productivity (linear correlation) than to map separate levels of overtime to distinct productivity levels (association rule mining). In addition, LCD can guide association rule discovery. For example, a relationship between factory region and productivity might suggest that association rule mining could pinpoint specific high productivity regions. Furthermore, even when no correlation exists, association rule mining can often generate artifacts or incorrect results [13]. Finally, linear correlation is a necessary precondition to infer causality [15]. Note that the above does not imply that linear correlation discovery is better: linear correlation discovery and association rule mining discover different things.

There are two measures of interest in a correlation: *strength* and *significance*. The *strength* of a correlation ranges from 0 to 1. This measures the amount of common dispersion between two attribute groups. The *significance* of the correlation, which also ranges from 0 to 1, is the probability that the correlation is coincidental. A lower score implies a genuine (i.e., non-coincidental) relationship. Three factors influence *significance*: (1) the strength of the correlation: the stronger the correlation, the more believable it is; (2) the sample size: the more times we see the correlation occur, the more believable it is; and (3) the level of skepticism (i.e., statistical power): the more skeptical the analyst is, the more evidence is needed to make the correlation believable [16].

2.2. Linear correlation discovery and data mining

In this research, we adopt statistical techniques (e.g., correlation measurement functions), but follow the data mining approach [17–19]. In the data mining approach, relationships are discovered in pre-existing data [7,20,21]. In contrast, in the hypothetico-deductive approach of classic inferential statistics, data is collected to test hypotheses proposed by humans [15]. Fig. 1 contrasts the two approaches. In the hypothetico-deductive approach, the *hypothesis* that overtime leads to

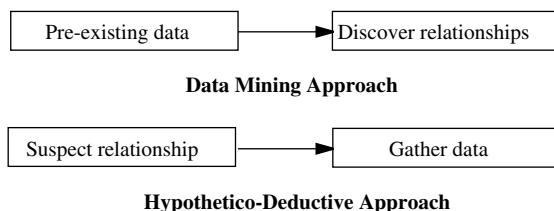


Fig. 1. Contrasting the data mining and hypothetico-deductive approaches.

Table 2
Definitions

| Term | Definition |
|------------------------------|--|
| Attribute | A column in a table (e.g., a relation in a relational database). The term is similar to “item” or “variable” in quantitative research, except that items and variables are theory laden |
| Attribute group | A set of attributes. The term used here is similar to the term “formative construct,” as used in quantitative research, except that theory is induced from the attribute group to the construct instead of from the construct to the attribute group |
| Confirmation | The demonstration that a relationship exists in a database. The term contrasts with validity, which indicates that a relationship is true in the real world |
| Correlation | A standardized measure of the covariance between two attribute groups |
| Domain class | The classification of the measurement properties of attributes and attribute groups |
| Linear | Achieving independence. That is, attributes in an attribute group are assumed to be uncorrelated, and interaction effects are not hypothesized |
| Linear correlation discovery | The automatic application of linear correlation functions to induce relationships from data |
| Relation | A ‘table’ in a relational database |
| Validity | The demonstration that a relationship is true in the real world |

lower performance drives the development of an experiment to test whether the hypothesis is true. In data mining, pre-existing overtime and productivity data lead to the *conclusion* that overtime is related to decreased performance.

The fundamental tradeoff between these two approaches is accuracy versus cost [22]. In the hypothetico-deductive approach, data is collected to rule out any threats to validity, and the data analyst has confidence that certain counter explanations are not plausible. For example, the analyst might rule out that overtime is caused by low productivity (the reverse hypothesis) because overtime was directly manipulated in the experiment. However, it is significantly less expensive to infer knowledge from the existing factory databases than to manipulate worker overtime.

Because this research leverages the statistics, data mining, and database literature, many terms herein have similar but not identical meanings to those that have been used in these research streams. It is therefore necessary to clearly define these commonly used terms, which are summarized in Table 2.

2.3. Linear correlation discovery process

Through our proposed methodology, we establish a five-step LCD process, as depicted in Fig. 2.

Step 1: Determine discovery parameters. The sample size to be analyzed and the thresholds for “interestingness” are determined and specified. In Section 3, we present the method to determine sample size. The thresholds for each step of the methodology are specified for the corresponding methods.

Step 2: Classify attributes. The relation (i.e., table) or query result for LCD is identified and extracted from the target database. It is assumed that the relation or query result to be analyzed has been identified a priori. The schema information of the imported relation

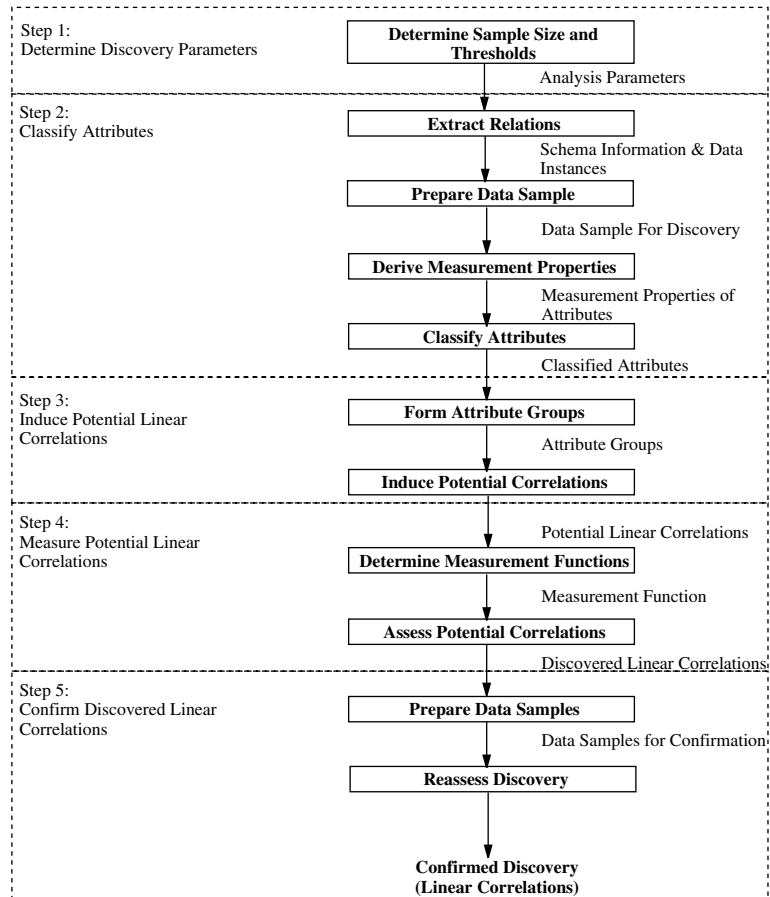


Fig. 2. Linear correlation discovery process.

is also obtained from the data dictionary of the target database and used to derive the attributes' measurement properties (e.g., distinctness, order and distance). In turn, measurement properties are used to classify the attributes and determine the appropriate correlation measurement functions to discover the linear correlations. In Section 4, we present a method for deriving these measurement properties by examining not only the schema information but also data instances. Because many widely accepted database management system (DBMS) interchange formats (e.g., ODBC, JDBC) exist, we do not discuss how relations are extracted and prepared for discovery.

Step 3: Induce potential correlations. When the attributes are classified, they form valid attribute groups according to their measurement properties. The resulting attribute groups are paired to highlight the potential linear correlations. We provide details about the generation of attribute groups and the induction of potential linear correlations in Section 4.

Step 4: Measure potential correlations. Beginning with the pairs with the fewest attributes, the potential correlations are then measured. We elaborate on this step in Section 5.

Step 5: Confirm discovered correlations. A set of data samples is extracted using the same sampling method as described in Step 2, to analyze and confirm the discovered linear correlations. The discovered correlations are then remeasured with the new samples. If the cumulative results surpass the threshold values, the discovered correlations are confirmed, as we discuss in Section 6.

3. Data sampling method

To determine the sample size for analysis, we apply Eq. (1), which estimates the sample size (N) on the basis of population proportions [23,24]. This formula does not require knowledge about the effect size or the shape of the population distribution. Such statistics are difficult to determine and obviate the need for sample analyses. Because the formula assumes very little about the properties of the population analyzed, it tends to overestimate the sample size required.

$$N = \frac{p \times (1 - p) \times Z(\alpha/2)^2}{\epsilon^2} \quad (1)$$

In Eq. (1), p is a value, between 0 and 0.5, that indicates a degree of variability; ϵ is the degree of error that varies between 0 and 1 (i.e., the precision of the sample); α is the probability of exceeding this degree of error (i.e., the accuracy of the sample); and Z is a function that describes the area under the standard normal curve [23,24]. If p , the degree of variability, is not known, the worst case value of 0.5 can be adopted. Because the formula assumes an infinite population, it can estimate a sample size for any relation (i.e., table) with many data instances (i.e., tuples). Finally, the thresholds α and ϵ are normally set between 0.01 and 0.05 [23,24].

Suppose that the distribution of values of the sample differs from the distribution of values in the population by, at most, 2.5% (ϵ), 95% of the time ($1 - \alpha$). Eq. (1) will then indicate that 1537 samples are sufficient (the Z score for $\alpha/2$, 0.025, is 1.96, thus $N = \frac{0.25 \times 1.96^2}{0.025^2} = 1536.64 \approx 1537$). Therefore, we would employ 1537 data samples for correlation induction and measurement, and discovery confirmation.

4. Induction of potential linear correlations

Potential linear correlations are induced in the following sequence: (1) *attribute classification*, (2) *attribute group generation*, and (3) *induction of potential linear correlations for discovery*. By identifying the set of measurement properties, we create the foundation for automatic attribute classification. In addition, these properties establish the attribute domain hierarchy depicted in Fig. 3. The three main classes of the attribute domain hierarchy, NOMINAL, ORDINAL and INTERVAL map to properties frequently employed in statistical analyses [25].

For example, the property *distinctness* occurs within the NOMINAL domain class. An attribute is distinct if two unique instances have different meanings, such as the values “China” and “Japan” in the attribute Country. The NOMINAL domain contains two terminal classes: CATEGORICAL, and DICHOTOMOUS. If an attribute possesses only two distinct instances, it

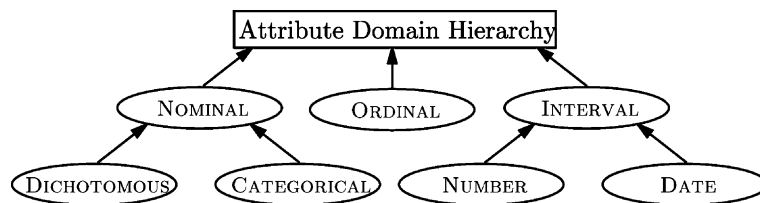


Fig. 3. Attribute domain hierarchy.

belongs to the **DICHOTOMOUS** domain class, in which attributes are subject to a limited distance comparison. For example, the attribute `Has_ERP` has only two possible values: Yes and No, and therefore belongs to the **DICHOTOMOUS** domain class; one can claim that $|Yes - No| = |No - Yes|$. Some correlation functions (e.g., logistic regression) exploit this property. In contrast, the **CATEGORICAL** domain class does not have a distance comparison property. Thus, the sample attribute `Factory_Location` belongs to the **CATEGORICAL** domain class; it is impossible to claim $|China - Japan| = |Japan - France|$.

The **ORDINAL** domain class includes both *distinctness* and *order* properties. The values of an attribute in this domain class are ranked. For example, the attribute values of `Transport_Service` are ranked from 1, which indicates a dependent and therefore dependable company, to 3, which indicates an unreliable company.

The **INTERVAL** domain class includes the properties of the **ORDINAL** class, as well as *distance*. It is therefore possible to determine the interval between two instances of this domain class. For example, with `Transport_Service`, it is not possible to determine how much better a dependent company is than an unreliable company. However, with `Avg_Wage`, a wage of \$4000 is better than \$3000 by \$1000. The **INTERVAL** domain class has two subclasses, **DATE** and **NUMBER**, containing attributes whose values indicate dates (e.g., `Start_Date`) or are subject to all arithmetic operations (e.g., `Avg_Wage`), respectively.

4.1. Attribute classification

The data type of an attribute can possibly map to more than one attribute classification. For example, the Integer data type does not necessarily imply a **NUMBER**. An Integer can represent a **DATE** (e.g., a date with the format YYYYMMDD), an **ORDINAL** (e.g., `Transport_Service`, where '1' means a dependent company, and '3' means an unreliable company), a **CATEGORICAL** (e.g., `Region`, where 1 is North America, 2 is Europe, and so forth), or a **DICHOTOMOUS** (e.g., `Has_ERP` where 1 is Yes) domain class. In Table 3, we describe the possible domain classes for each data type. Because of the variety of potential domain classes, the proposed LCD methodology must analyze not only the schema information (including its data type, attribute length, data format and input masks), but also data instances. On the basis of those derived measurement properties, heuristic rules then determine and assign the most appropriate domain class to each attribute.

Although many are possible, only one class from the attribute domain hierarchy {**DICHOTOMOUS**, **DATE**, **ORDINAL**, **CATEGORICAL**, **NUMBER**} can be assigned to an attribute. Some attributes may not fit into any domain class; these would not be considered for LCD. We present a

Table 3
Possible domain classes of each data type

| Data type | NUMBER | DATE | ORDINAL | CATEGORICAL | DICHOTOMOUS |
|-----------|--------|------|---------|-------------|-------------|
| Integer | ✓ | ✓ | ✓ | ✓ | ✓ |
| Decimal | ✓ | ✓ | | | |
| String | ✓ | ✓ | ✓ | ✓ | ✓ |
| Date | | ✓ | | | |

thorough discussion of these heuristic rules in [Appendix A](#), but here, we briefly illustrate their use for classifying the *Region* attribute with distinct instances (North America, Europe, Asia, South America, and Africa). *Region* has a String data type. Furthermore, because its instances contain alphabetic characters, it cannot be NUMBER. Although the number of distinct instances in the attribute are fairly small, there are more than two, so it cannot be DICHOTOMOUS. It cannot be DATE, because the instances do not contain a month sequence (e.g., January, February, March). Finally, it cannot be ORDINAL, because the variation in the length of the instances is very high and the first letters of the sorted values do not cover the entire alphabet. That is, the sequence does not have an instance that begins with the letter “B” instead, “(A)sia”, is followed by “(E)urope,” which means that all the letters from “B” to “D” are skipped. Therefore, the only domain class that *Region* can be assigned to is CATEGORICAL.

4.2. Attribute group generation

After attributes have been classified according to the domain hierarchy, the proposed LCD methodology generates valid attribute groups. A set of attributes is considered as a valid group if a proper domain class can be assigned to the entire group. By removing any attribute groups without a proper domain class assigned, the proposed LCD methodology reduces the number of attribute groups to be considered and expedites the discovery process.

The generation of attribute groups proceeds as follows: first, each attribute is treated as a single-attribute group. Second, two attributes are combined if their domain classes result in a valid domain class, according to [Table 4](#). Third, the attribute groups with two attributes are then combined with another attribute to form three-attribute groups. Again, [Table 4](#) provides the means to determine whether the resultant attribute group has a valid domain class assigned.

Table 4
Summary of domain class assignments

| Attr. Grp B | Attr. Grp A | | | | |
|-------------|-------------|------|-------------|-------------|-------------|
| | NUMBER | DATE | ORDINAL | CATEGORICAL | DICHOTOMOUS |
| NUMBER | NUMBER | DATE | N/A | N/A | NUMBER |
| DATE | DATE | DATE | N/A | N/A | DATE |
| ORDINAL | N/A | N/A | CATEGORICAL | CATEGORICAL | CATEGORICAL |
| CATEGORICAL | N/A | N/A | CATEGORICAL | CATEGORICAL | CATEGORICAL |
| DICHOTOMOUS | NUMBER | DATE | CATEGORICAL | CATEGORICAL | DICHOTOMOUS |

In most cases, an attribute group is assigned to the domain class with the fewest properties, which ensures that subsequent analysis does not employ a property that is inapplicable to some attributes. For example, when an ORDINAL attribute and a CATEGORICAL attribute are combined, the result is a CATEGORICAL attribute group. There are exceptions to this rule. For example, the combination of NUMBER or DATE with CATEGORICAL would produce too many instances to be analyzed meaningfully as CATEGORICAL, and therefore, such cases are discarded. However, when a NUMBER attribute group is combined with a DICHOTOMOUS attribute group, the resultant attribute group can be assigned as a NUMBER attribute group because the DICHOTOMOUS attribute group possesses the distance property.

Because Table 4 is associative, the order of combination does not affect the resulting domain class. For example, the attributes Region, Transport_Service and Has_ERP, are assigned the CATEGORICAL, ORDINAL and DICHOTOMOUS domain class, respectively. According to Table 4, the resultant attribute group {Region, Transportation_Service, Has_ERP} is assigned the CATEGORICAL domain class regardless of the sequence in which the three attributes were combined into an attribute group.

The same domain class assignment procedure can be applied to attribute groups with four or more attributes. However, in the evaluation experiments, we restrict the size of the attribute groups to three. There are two reasons for this limit. First, the smaller size reduces the computational effort of the discovery process. Second, correlations between large attribute groups typically are caused by correlations between smaller attribute groups [11]. In turn, it is usually unnecessary to automate the linear correlation of large attribute groups; such correlations can be derived instead through visual inspection.

After the attribute groups are formed, the pair of any two attribute groups that have no attributes in common is a candidate for linear correlation discovery. For example, the attribute groups {Region, Start_Date} and {Defects} might be paired, but {Region, Start_Date} and {Region} cannot be.

5. Measurement of potential linear correlations

5.1. Determination of measurement functions

For each attribute group pair, an appropriate measurement function must be determined to evaluate its linear correlation (Table 5). The appropriate measurement function is determined on the basis of the domain classes of the attribute groups to be correlated. For example, the Box–Cox function is appropriate for measuring the correlation of the pair {{Avg_Wage}, {Widgets_Made}} because both attributes belong to the NUMBER domain class. If we pair the attribute group {Cost_Unit, Widgets_Made} (NUMBER domain class) with the attribute group {Transport_Service, Country} (CATEGORICAL domain class), according to Table 5, MANOVA is the appropriate correlation measurement function.

We choose 12 measurement functions (summarized in Table 5), which represent an extension of the attribute characteristic/measurement function mapping framework found in [26] (p. 269). Each chosen correlation function can be applied only to attributes with specific measurement properties (e.g., distinctiveness and order). We do not elaborate on these measurement functions

Table 5
Selection of correlation measurement functions

| Group A | Group B | | | | | | |
|---------------------|---------------------|------------|-------------------|----------|-----------|-----------|-----------|
| | NUMBER (<i>N</i>) | NUMBER (1) | DATE (<i>N</i>) | DATE (1) | ORD. | CAT. | DICH. |
| NUMBER (<i>N</i>) | CC | BT | CC | BT | OL | MA | LO |
| NUMBER (1) | BT | BC | BT | BC | OL | AN | PB |
| DATE (<i>N</i>) | CC | BT | CC | R^2 | OL | MA | LO |
| DATE (1) | BT | BC | R^2 | R^2 | OL | AN | PB |
| ORD. | OL | OL | OL | OL | ρ | λ | λ |
| CAT. | MA | AN | MA | AN | λ | λ | λ |
| DICH. | LO | PB | LO | PB | λ | λ | ϕ |

BC = Box–Cox.

BT = Box-Tidwell.

R^2 = Pearson's coefficient of determination.

CC = Canonical correlation.

OL = Ordered logit.

ρ = Spearman's coefficient of rank determination.

AN = One-way analysis of variance (ANOVA).

MA = One-way multivariate analysis of variance (MANOVA).

λ = Goodman and Kruskal's lambda.

LO = Logistic regression/ANOVA/MANOVA.

PB = Point biserial correlation coefficient.

ϕ = Phi coefficient.

here because they are commonly discussed in most statistics textbooks (e.g., [11,27,28]). However, note that the attribute groups of the INTERVAL domain class are subcategorized into single (1) and multi-*(N)* attribute groups to reflect the variation present in many such correlation functions. This distinction does not appear for other domain classes. Multi-attribute groups cannot be assigned to the ORDINAL domain class (see Section 4.2), and those assigned to the CATEGORICAL domain class can be treated as if they were single attributes. For example, the attribute group {Transport_Service, Has_ERP} can be treated as if it contains four instances, (Yes, Yes), (Yes, No), (No, Yes), and (No, No). Finally, multi-attribute groups assigned to the DICHOTOMOUS domain class are treated as multi-attribute NUMBERS, in that the DICHOTOMOUS domain class contains the limited distance semantic. We explain our rationale for selecting these functions as optimal for LCD in Appendix B.

5.2. Assessment of correlations

The correlation measurement function deemed appropriate for a pair of attribute groups (X, Y) is then evaluated to obtain the strength $S_{(X,Y)}$ and significance $p_{(X,Y)}$ scores. As defined previously, the strength score measures the proportion of explained variance, or the strength of the correlation between the attribute groups (e.g., R^2), whereas the significance score measures the probability that the covariance occurs by chance. An attribute group pair (X, Y) is considered a discovered linear correlation if its strength score $S_{(X,Y)}$ is greater than a strength threshold γ , and the statistical significance $p_{(X,Y)}$ is less than a threshold α . The commonly chosen default threshold for α is

0.05 [29], and 0.1, 0.3, and 0.5 (i.e., R^2 values of 0.01, 0.09 and 0.25, respectively) represent good small, medium, and large threshold values of γ [16]. Then, to detect and reject artifact linear correlations automatically, the discovered linear correlations are subjected to further analysis with additional data samples.

For the pair (X, Y) with a linear correlation, any attribute group that contains this pair as a subset is flagged as non-analyzable, because the pair's strong correlation will distort any correlation computed on supersets of that pair. For example, if the correlation between $\{\text{Worker_Age}\}$ and $\{\text{Avg_Wage}\}$ are strong and significant, the correlation between $\{\text{Worker_Age, Avg_Wage}\}$ and $\{\text{Defects}\}$ will be distorted and therefore should not be considered.

6. Confirmation of discovery

By developing a discovery confirmation method, we aim to eliminate artifact discoveries from the automatic LCD process. This method adopts a common practice for validating theories in science [30], in that it repeats the discovery measurement step on a set of new data samples drawn from the original relation. This repetition serves to increase our confidence in the discovered linear correlations.

Consider, for example, that a scientist discovers that smoking and cancer are correlated with a p -value of 0.03. Three associates of the scientist independently replicate the experiment on new samples and obtain separate p -values of 0.03, 0.04, and 0.05. A meta-analytical test (e.g., Fisher's Combined Test: see Eq. (2)), shows that the probability that the three scientists could all obtain such low p -values is substantially less than 0.03. We therefore can establish from their combined measures that a correlation exists between smoking and cancer. The confirmation method for our proposed LCD methodology uses the same principle and proceeds with the following steps:

- (1) *Specify threshold values.* Three threshold values must be specified for the confirmation: (1) the all-pairs significance threshold value α'_R , (2) the all-pairs effect size threshold σ , and (3) the all-pairs power threshold \mathbb{B} . An *intermediate significance threshold value* then should be calculated from the all-pairs significance threshold value using a traditional technique that adjusts for alpha inflation (e.g., Bonferroni, Tukey-w, Dunn, Dunn-Šidák). The Bonferroni technique, a conservative approach in which $\alpha = \frac{\alpha'_R}{D}$, where D is the number of attribute group pairs to test can be adopted as the default [31,32].
- (2) *Generate confirmation data samples.* Initially, random sampling of the original relation generates two data sets. The first set controls for artificial correlations discovered in the initial analysis. The second set controls for artificial correlations discovered in the first test set. To confirm a discovered correlation, more test data sets may be generated.
- (3) *Repeat measurement functions.* We obtain the strength and significance scores of the confirmation data sets by executing the originally chosen correlation measurement function.
- (4) *Compare the obtained p -values.* In the example of the scientist who found a correlation between smoking and cancer, the actual confidence of the statistical significance score for the replicated results could be established using a meta-analytical test such as the Fisher, Winer or Stouffer test. For discovery confirmation, Fisher's combined test [33] is also adopted

to confirm the statistical significance of the test set. Eq. (2) represents Fisher's combined test, where k is the number of resamples, p_i is the p -value obtained for the correlation from each test set, and $\chi_{df=2k}^2$ is the chi-square statistic with $2k$ degrees of freedom.

$$p' = \chi^2 \left(-2 \sum_{i=1}^k (\ln p_i), df = 2k \right) \quad (2)$$

$\chi^2(a, b)$ indicates the p -value of the chi-square distribution given score a and degrees of freedom b . When this value (p') is compared with the intermediate significance threshold value, if the threshold value is greater, the correlation is confirmed. We prefer Fisher's combined test to other meta-analytical tests (e.g., Winer and Stouffer) because it has been shown to be more robust [23,28,34].

- (5) *Derive the effect size for each pair of attribute groups.* Using Eq. (3), we determine the effect size (S), the strength of the correlation from test set i (R_i^2), and the number of test sets (n) [33,16,35]. The meta-analytic p -value and the effect size enable us to estimate both the power of the test set and the interestingness of the correlation.

$$S \approx \sqrt{\frac{\sum R_i^2}{n}} \quad (3)$$

- (6) *Derive the individual power for each test set and discovered correlation.* With Eq. (4), we can determine the individual power for each test set, where $Z(x)$ is the score of x on the Z -distribution (standard normal curve); b_i is the statistical power for test set i ; α and S are the statistical significance threshold and effect size for the correlation, respectively; and n is the number of tuples in the test set [16].

$$Z(b_i) = \left(\operatorname{arctanh} S + \frac{S}{2(n-1)} \right) \sqrt{n-3} - Z(1-\alpha) \quad (4)$$

- (7) *Derive the individual power for each discovered correlation.* For the application of the Fisher Combined Test (Eq. (2)), where b_i is substituted for p_i , and β for p' , if the individual power of the correlation is below the threshold value \mathbb{B} , a new test set must be obtained, and the confirmation method reiterates from Step 2.
- (8) *Compare the all-pairs significance and all-pairs effect size with threshold values.* If the significance threshold of all-pairs is greater than the obtained meta-analytic score and the effect size of all-pairs is less than the meta-analytic score, the correlation is confirmed. Otherwise, it is rejected.

This confirmation method can be confounded when the individual samples have low power (i.e., <0.50). To control this situation, resampling should be performed 15 times at most (i.e., the initial 2 test sets and up to 13 iterations). Any confirmation that requires more than 15 samples should be considered inconclusive. We use 15 as the threshold because Fisher's Combined Test, in which each sample contributes 2 degrees of freedom, is based on a chi-square distribution that does not change substantially after 30 degrees of freedom [28,29,34].

The factory example falls within this limit with its twelve attributes for knowledge discovery (Region, Country, Worker_Age, Avg_Overtime, Avg_Wage, Start_Date, Transport_Service, Has_ERP, Cost_Unit, Widgets_Made, Defects, and Factory_Downtime). If we consider only single-attribute linear correlation discovery, we have 144 possible correlations (i.e., 12×12). However, only some of the relationships are of interest; for example, we are not interested in the relationship between Country and Start_Date on the basis of their measurement properties. Specifically, we are interested in the relationships between the eight descriptive attributes: Region, Country, Worker_Age, Avg_Overtime, Avg_Wage, Start_Date, Transport_Service, and Has_ERP and four result-oriented attributes: Cost_Unit, Widgets_Made, Defects, and Factory_Downtime. The data analyst is responsible for ascertaining the useful predictor (e.g., Avg_Wage) and outcome (e.g., Cost_Unit) attributes. Therefore, in the example, the LCD process should consider only 32 (i.e., 8×4) possible single-attribute relationships. If we set the significance threshold at 0.05 (i.e., $\alpha = 0.05$), the accepted intermediate significance threshold for one pair will be $0.05 \div 32 = 0.0015$. The statistical power and effect size thresholds are 0.8 and 0.06, respectively [16,29].

Assume that the initial measurement of the correlation between Avg_Overtime and Defects has a strength of 0.062 (see Table 6). This strength score suggests a correlation between Avg_Overtime and Defects and thereby requires further confirmation. The confirmation method resamples two new data sets, trial 1 and 2, from the database, and finds significance scores of 0.0289 and 0.057 and strength scores of 0.06 and 0.05, respectively. The overall significance and strength scores for trial 1 are thus 0.0122 and 0.055. Because the combined power score is less than 0.80, the confirmation process iterates, and each additional trial generates a new confirmation data sample. At the sixth trial, a satisfactory level of power has been achieved (i.e., $0.839 > 0.8$), which means that the combined statistical significance of the seven test samples is less than 0.0015, and the discovered correlation is confirmed; a correlation really exists between Avg_Overtime and Defects. However, because the strength (effect size) of this pair is less than the threshold value (i.e., $0.055 < 0.06$), the correlation should be rejected as uninteresting.

Table 6
Example of discovery confirmation $N = 1000$

| Trial | Trial set # | Significance | | Strength | | Power | |
|-----------|-------------|---------------|----------|--------------|----------|---------------|----------|
| | | Test set | Combined | Test set | Combined | Test set | Combined |
| Discovery | | 0.0269 | | 0.062 | | 0.6127 | |
| 1 | 1 | 0.0289 | 0.0289 | 0.060 | 0.060 | 0.6005 | 0.6005 |
| 1 | 2 | 0.0570 | 0.0122 | 0.050 | 0.055 | 0.4750 | 0.6430 |
| 2 | 3 | 0.1031 | 0.0081 | 0.040 | 0.050 | 0.3521 | 0.5965 |
| 3 | 4 | 0.0411 | 0.0025 | 0.054 | 0.051 | 0.5382 | 0.6656 |
| 4 | 5 | 0.0199 | 0.0005 | 0.066 | 0.054 | 0.6604 | 0.7566 |
| 5 | 6 | 0.0358 | 0.0001 | 0.060 | 0.055 | 0.5633 | 0.7995 |
| 6 | 7 | 0.0311 | <0.0001 | 0.055 | 0.055 | 0.5882 | 0.8390 |

7. Evaluation experiments

We have developed a prototype system, the Linear Correlation Discovery System, to evaluate the proposed LCD methodology [10]. In a first set of experiments, we evaluate the effectiveness of the heuristic rules for attribute classification. We also perform a second set of experiments to evaluate the effectiveness of the LCD methodology.

7.1. Heuristic rules for attribute classification

To assess the effectiveness of the heuristic rules, we compare the domain classes derived by the heuristic rules for attribute classification with the given semantic descriptions provided for the public domain data sets [36–47], most of which were obtained from the CMU, UCI, and World Bank data repositories [48–50]. Of a total of 319 attributes, the domain classes of 307 attributes were assigned correctly, which indicates an accuracy rate of 96.2% for the heuristic rules. Only one of the data sets had more than one error per 10 attributes.

Table 7
Experimental results on heuristic rules for attribute classification

| Data set | # Attr | # Attr√ | Incorrect classification | Reason for failure |
|-----------------------------|--------|---------|---|---|
| [36] | 9 | 7 | No_of_Cylinders and Time_to_accelerate were classified as ORDINAL instead of NUMBER | Too few distinct instances |
| [37] | 11 | 10 | Education was classified as ORDINAL instead of INTERVAL | Too few distinct instances |
| [38] | 8 | 7 | Sample_Date was classified as NUMBER instead of as DATE | Day values of the attribute were given the value '00' |
| [39] | 33 | 33 | No error | No error |
| [40] | 6 | 6 | No error | No error |
| [41] | 6 | 6 | No error | No error |
| [42] | 15 | 14 | Education was classified as CATEGORICAL instead of ORDINAL | Instances of EDUCATION were found in the dictionary |
| [43] | 27 | 25 | #_in_family and #_kids classified as ORDINAL instead of NUMBER | Too few distinct instances |
| [44] | 4 | 4 | No error | |
| [45] ^a (City) | 25 | 23 | stfips and plfips classified as NUMBER instead of CATEGORICAL | Both of these are city region codes. However, there are over 30 possible distinct codes |
| [45] ^a (Metro) | 29 | 28 | MSACode identified as NUMBER instead of CATEGORICAL | There are over 30 distinct MSA codes |
| [45] ^a (Country) | 32 | 30 | stfips and plfips classified as NUMBER instead of CATEGORICAL | Both of these are city region codes. However, there are over 30 possible distinct codes |
| [46] ^b (6095) | 77 | 77 | No error | |
| [46] ^b (Panel) | 30 | 30 | No error | |
| [47] | 7 | 7 | No error | |

^a This data collection contains three separate data sets.

^b This data collection contains six separate data sets. Only two of the data sets come with accompanying descriptions.

In Table 7, which summarizes the experimental result, **# Attr** indicates the number of attributes in a data set; **# Attr_✓** indicates the number of attributes correctly assigned by the heuristic rules; **Incorrect Classification** describes those attributes assigned to incorrect domain classes, as well as the nature of the error; and **Reason for Failure** summarizes the reasons behind the incorrect assignment. Half of the incorrectly assigned attributes should have been classified as **NUMBER** but instead were classified as **ORDINAL**. These attributes have the following common syntactic features:

- The data values are represented by Integers and have few distinct values. The combination of these two characteristics caused them to be misidentified as **ORDINAL**.
- They are predominantly used to count something (e.g., years of education, number of children). The sole exception was the attribute “Time to Accelerate.”

In consideration of these characteristics of misclassified attributes, we intend to extend and revise the heuristic rules and perform additional experiments in the future to further improve their accuracy.

7.2. Experiments on the effectiveness of linear correlation discovery

In the second set of experiments, we compare the LCD methodology with the findings of human experts. As surrogates for human experts, we used research papers that analyzed the data sets used in Section 7.1. Three data sets from the classification experiments were chosen [40,45,47]. For the other data sets, we either lacked sufficient expertise to understand or could not obtain the related papers.

The analysis was performed using the Linear Correlation Discovery System described in [10]. Each run took from 30min to 1h. Delays in execution were primarily associated with the SPSS batch processor, which would load and shut down the SPSS server for each batch. The implementation was written so that each analysis was one batch job.

In Tables 8–10, we summarize the data sets and compare the findings of the original researchers with ours. Because the LCD methodology depends partly on user-defined thresholds, we set three threshold levels for strength: strong (0.09), moderate (0.04), and weak (0.01). We also established

Table 8
Comparison of methodology versus published research: [40]

| Original findings | Threshold | | | | |
|--|-----------|-------|--------|----------|------|
| | Str. | Sig. | Strong | Moderate | Weak |
| Father's occupation predicts status of first job | 0.060 | 0.000 | No | Yes | Yes |
| Father's occupation predicts status of current job | 0.075 | 0.000 | No | Yes | Yes |
| Son's occupation predicts status of first job | 0.068 | 0.000 | No | Yes | Yes |
| Son's occupation predicts status of current job | 0.077 | 0.000 | No | Yes | Yes |
| Race predicts status of first job | 0.097 | 0.000 | Yes | Yes | Yes |
| Race predicts status of current job | 0.135 | 0.000 | Yes | Yes | Yes |
| Family disruption predicts status of first job | 0.076 | 0.000 | No | Yes | Yes |
| Family disruption predicts status of current job | 0.106 | 0.000 | Yes | Yes | Yes |

Table 9
Comparison of methodology versus published research: [45]

| Data set | Original findings | Threshold | | | | |
|-----------|--|-----------|-------|--------|----------|------|
| | | Str. | Sig. | Strong | Moderate | Weak |
| (City) | Ethnic heterogeneity affects spending on roads | 0.066 | 0.000 | No | Yes | Yes |
| | Ethnic heterogeneity affects spending on sewage and trash pickup | 0.009 | 0.000 | No | No | No |
| | Ethnic heterogeneity affects spending on police | 0.022 | 0.000 | Yes | Yes | Yes |
| | Ethnic heterogeneity affects spending on fire protection | 0.000 | 0.789 | No | No | No |
| | Ethnic heterogeneity affects spending on roads per capita | 0.015 | 0.000 | No | No | Yes |
| (Metro) | Ethnic heterogeneity affects spending on roads | 0.224 | 0.000 | Yes | Yes | Yes |
| | Ethnic heterogeneity affects spending on police | 0.063 | 0.000 | No | Yes | Yes |
| | Ethnic heterogeneity affects spending on education | 0.063 | 0.000 | No | Yes | Yes |
| | Ethnic heterogeneity affects spending on health | 0.084 | 0.000 | No | Yes | Yes |
| | Ethnic heterogeneity affects spending on roads per capita | 0.139 | 0.000 | Yes | Yes | Yes |
| (Country) | Ethnic heterogeneity affects spending on roads | 0.157 | 0.000 | Yes | Yes | Yes |
| | Ethnic heterogeneity affects spending on police | 0.099 | 0.000 | Yes | Yes | Yes |
| | Ethnic heterogeneity affects spending on education | 0.029 | 0.000 | No | No | Yes |
| | Ethnic heterogeneity affects spending on health | 0.034 | 0.000 | No | No | Yes |
| | Ethnic heterogeneity affects spending on welfare | 0.020 | 0.000 | No | No | Yes |
| | Ethnic heterogeneity affects spending on roads per capita | 0.098 | 0.000 | Yes | Yes | Yes |

Table 10
Comparison of methodology versus published research: [47]

| Original findings | Threshold | | | | |
|---|-----------|-------|--------|----------|------|
| | Str. | Sig. | Strong | Moderate | Weak |
| Equipment investment affects GDP growth | 0.226 | 0.000 | Yes | Yes | Yes |
| Non-Equipment investment affects GDP growth | 0.344 | 0.000 | Yes | Yes | Yes |

the significance thresholds at 0.05. The result was required to be greater than the strength threshold and less than the significance threshold to be considered interesting.

In the study involving the “Effects of Family Disruption on Social Mobility” [40] data set, the authors used a sophisticated analytic model (SAT) to ascertain that a person’s father’s occupation, (eldest) son’s occupation, race, and family disruption level (e.g., separated parents, living with grandparents) correlated with the prestige level of the person’s first and current jobs. As Table 8 indicates, the proposed LCD methodology replicated the authors’ findings when a moderate effect size was used.

In the “Public Goods and Ethnic Divisions” [45] study, the authors used traditional measures of linear correlation, identical to those used herein, to demonstrate that ethnic heterogeneity, defined as environments in which the most populous ethnic race was only marginally more numerous than the next most populous race, caused governments to reallocate tax monies away from programs that provided unequal benefits to different races and toward programs that all races would value equally. For example, the authors found that ethnic heterogeneity was correlated

with reduced spending on schools (some races value formal education more than others) and increased spending on hospitals (serious illnesses must be treated regardless of race). The study employed data sets with three levels of detail: city, metropolitan area, and country. Our findings are identical to the published results in the paper “Public Goods and Ethnic Divisions” [45] (Table 9). However, our interpretation of the results differs from that of the authors in several ways. For example, they conclude that ethnic heterogeneity affects government expenditure on fire departments. The t -values for this attribute in the original study was 0.002; Information systems (IS) research uses thresholds for t -values that are based on statistical significance at the 0.05 level. Thus, an IS researcher would have accepted this conclusion only if the t -value was at least 1.65 [25].²

In the final study, “How Strongly Do Developing Countries Benefit from Equipment Investment?” [47], the authors attempted to ascertain the relationship between investment in capital equipment and gross domestic growth (GDP). The LCD methodology was able to replicate this study’s results even at the strong threshold 10.

The results of the second set of experiments indicate that the proposed LCD methodology is useful and effective in automating the process of LCD in that our experimental results largely agree with those of the studies we replicated. The main difference between our results and those of the original authors lies in our interpretation of the results. In addition, in many cases, the original authors elected to accept thresholds that are substantially lower than those we recommend, which is understandable given our substantively distinct perspectives. The original authors, following the hypothetico-deductive approach, were attempting to demonstrate the existence of a phenomenon. Thus, they needed to test for only a few things, and it was of greater importance to demonstrate the existence of a phenomenon than to investigate its strength. In contrast, the LCD methodology follows an exploratory, or data-mining approach, in which numerous possibilities (linear correlations) are tested. Furthermore, because it is unreasonable to expect that every discovered phenomenon can be acted on, only the strongest phenomena are of real importance, whereas weak phenomena are discarded as uninteresting.

8. Conclusion and future research

We present an automatic discovery methodology to expedite LCD from relational databases. The proposed LCD methodology improves knowledge discovery by enabling researchers to:

- (1) *Discover linear correlations automatically.* The methodology automates LCD by inferring the measurement properties of attributes, which are derived by examining schema information and the attributes’ values. Our established set of heuristic rules can determine the proper correlation measurement function for each potential linear correlation.
- (2) *Consider attribute groups when identifying correlations.* By identifying potential correlations, the methodology considers correlations between sets of attributes (i.e., attribute groups), as well as between individual attributes.

² In the original study [45], the authors only report R^2 values for an overall model and derive conclusions about individual effects from the t -values.

- (3) *Confirm the discovered correlations.* Because the discovery confirmation method repeats the measurement of the discovered correlations on a set of data samples drawn from the original relation, artifact discoveries can be rejected automatically.

Our work on LCD has opened up several avenues of further research. For example, the LCD methodology considers only correlations of attribute groups with randomly distributed errors. It would be interesting to consider attribute groups with other error distributions as well. For example, time series analysis is performed on attribute groups in which errors are distributed according to a time sequence [12]. The automation of time series analysis remains an unsolved research problem.

The LCD methodology discovers correlations without considering the direction of their relationships. Whereas the LCD methodology can determine that `Defects` and `Avg_Wage` are related, it cannot determine whether increasing `Avg_Wage` leads to lower `Defects` or if lower `Defects` leads to a better `Avg_Wage`. One extension might automatically establish structural equation models [51], or models of directional connections, that are based on the discovered correlations.

In addition, it is necessary to derive the measurement properties of attributes to automate LCD, but it is computationally intensive to derive these properties through the analysis of schema information combined with data instances. This computational effort might be reduced if the measurement properties could be captured during the database design process and then embedded into the database schema. Determining the measurement properties that might be appropriate for embedding is an interesting research topic for database design.

Our method also adopts fairly primitive methods to screen out uninteresting pairs of multi-attribute data. Other mechanisms to screen out pairs of data with low correlation (e.g., those in [52,53]) could potentially enhance our method's performance.

Finally, whereas database management systems have standard query languages and interfaces (e.g., SQL, ODBC) to facilitate easy access to data, such standardization does not exist for statistical data analysis packages. A language similar to SQL for data analysis should be developed to facilitate data analysis and knowledge discovery.

Acknowledgements

We thank Dr. Veda Storey for advice and comments, Dr. Amit Das, Dr. Michael Li, the UCLA Statistics Department (especially Dr. Jan de Leeuw, Dr. Richard Berk, and Dr. Henry Rubin), and the sci. stat. consult newsgroup community for their invaluable advice regarding the statistics in this paper, Elisabeth Nevins Caswell for writing assistance, as well as the multitude of people whose suggestions in one way or other have shaped this article.

Appendix A. Attribute classification

The following are the heuristic rules for assigning attributes to domain classes `DICHOTOMOUS`, `DATE`, `ORDINAL`, `CATEGORICAL`, and `NUMBER`.

A.1. Dichotomous

(1) The data type of the attribute is Integer or String and the attribute has only two distinct instances.

A.2. Date

An attribute has more than two distinct instances and satisfies one of the following conditions.

- (1) The attribute is of the Date data type.
- (2) A date-specific function (e.g., Month(), Day_of_Week()) is applied to the attribute in an application of the database system. Many database management systems incorporate interfaces to programming languages such as C or Visual Basic. The application source code is searched for reserved words associated with an attribute. For example, if “Month(Start_Date)” is found in the application source code, there is strong evidence that Start_Date should be assigned the DATE domain class.
- (3) The character strings “January”, “February” etc. are found in all instances.
- (4) The attribute values have a consistent format that allows only numeric character values and a single, non-numeric, non-alphabetic character. That character must be found in one of the following positions:
 - (a) Positions 3 and 6 (e.g., DD/MM/YY, MM-DD-YYYY, YY.MM.DD);
 - (b) Positions 5 and 8 (e.g., YYYY.MM.DD);
 - (c) Position 4 (e.g., DDD/YY);
 - (d) Positions 2 and 4, 2 and 5, 3 and 5, or 3 and 6 (e.g., in some date formats, January 1, 1999, January 25, 1999, December 1, 1999, and December 25, 1999 are stored as 1/1/99, 1/25/99, 12/1/99, and 12/25/99); or
 - (e) Positions 5 and 7, or 5 and 8 (e.g., 1999.1.1, 1999.25.1).
- (5) An attribute has the Decimal data type and satisfies the following conditions:
 - (a) It has an integer component of 5 or 7 digits and,
 - (b) Its values correspond to dates within a specified range. For five-character attributes, for example, the range is set at 15020.5 and 51909.5. For seven-character attributes, the range is set at 2415020.5 and 2451909.5.
- (6) An attribute has the String data type and satisfies the following conditions:
 - (a) It has a length of 5, 6, 7, or 8 characters (e.g., DDDYY, DDMMYY, DDDYYYY, DDMMYYYY);
 - (b) It has character values 0..9; and
 - (c) Its values conform to some accepted date representation system (e.g., the value ‘605074’ does not appear in any commonly accepted date representation system, but the value ‘123174’ [December 31, 1974] does).
- (7) An attribute has the Integer data type and satisfies the following conditions:
 - (a) It has 5-8 digits and
 - (b) Its values conform to some accepted date representation system (e.g., DDDYY, Julian, or Modified Julian date representation system).

A.3. Ordinal and categorical

Neither DATE nor DICHOTOMOUS can be assigned to the attribute and it satisfies one of the following conditions.

- (1) The attribute has a data length of less than 10 with less than 26 distinct values. The values 10 and 26 are defaults and can be adjusted according to the database being mined.
- (2) The attribute is a foreign key and is the candidate key of a small relation (fewer than 26 instances) or look-up table.

Justification: This rule attempts to capture CATEGORICAL or ORDINAL attributes that are presented as codes, such as departmental or occupational codes. Such codes are normally defined in a separate supporting table (called a look-up table). The value 26 is a default [2].

- (3) During data entry, the attribute values are selected from a list (e.g., pop-up window). Selection from a pop-up window can be discovered by searching for pop-up-related reserved words in the application source code.

The following four rules determine whether the attribute should be assigned the ORDINAL or CATEGORICAL domain class.

- (1) If an attribute has the Integer data type and a length greater than 2 (the size required for 25 distinct instances), it is assigned the ORDINAL domain class.

Justification: This rule captures ORDINAL attributes which have character positions with independent meaning. For example, the Mercedes-Benz E series is structured on a three digit code, where the first digit refers to the size of the car and the remaining digits refer to the model. Thus, a Mercedes-Benz E 200 is a smaller car than the Mercedes-Benz E 450.

- (2) If the values of the attribute can be meaningfully ranked, the attribute is assigned the ORDINAL domain class. For example, the values of the attribute Transport_Service can be ranked as (1, 2, 3). We consider an attribute meaningfully ranked if it satisfies one of the following conditions.
 - It is updated with the same sequence or pattern from a temporal point of view. For example, at universities, an “Assistant Professor” usually becomes an “Associate Professor” and then a “Professor”.
 - It contains non-numeric character strings. If alphabetically sorted, the left-most character of each value is at most two characters less than the next attribute value.

Justification: This rule enables LCD process to identify ORDINAL attributes such as Letter_Grade (i.e., A, B, C, D, and F), and Signal_Codes (i.e., Able, Baker, Charlie, Delta, Echo, and so forth).

- (3) If the values of an attribute can be found in a dictionary or spell-checking reference, the attribute is assigned the CATEGORICAL domain class. For example, countries (e.g., Afghanistan, Britain, China) will be found in a dictionary.

- (4) If all distinct instances of the attribute except one can be arranged in running order and the exception ends with a 9, the attribute is assigned the CATEGORICAL domain class. For example, an attribute with instances {1,2,3,4,5,6,7,8,9,10,99} would be assigned to the CATEGORICAL domain class.

Note: It is sometimes difficult to classify attributes as either ORDINAL or CATEGORICAL on the basis of schema information and instances. In such cases, the analyst must determine the appropriate domain class. When the analyst is unable to do so, the CATEGORICAL domain class is assigned, which ensures that the ordering semantic is not mistakenly assigned to an attribute.

A.4. Number

The NUMBER domain class is assigned to an attribute if it satisfies one of the following conditions.

- (1) The instances of the attribute are displayed with a measurement symbol (e.g., \$1200.00, 25°C, 35µm.).
- (2) The attribute has the Integer or Float data type, is not a candidate or foreign key, and is never displayed on a report or screen in a non-numerical format. For example, the attribute `Social_Security_Number` is displayed as 999-99-9999 and thus would not be assigned to the NUMBER domain class.
- (3) The attribute has the String data type, contains only numeric and associated characters (e.g., “.”, “-”), is not a candidate or foreign key, and is never displayed in a report or screen in a non-numerical format.

Appendix B. Justification for function assignment

B.1. Linear regression functions

An attribute group pair assigned the DATE domain class cannot be multiplied or divided. Thus, the correlation of any pair with the DATE domain class will always be linear with respect to the attribute groups. Consider the pair of attribute groups, {{`Start_Date`}, {`Defects`}}. The linear function that correlates them will be of the form $\text{Start_Date} + C = \text{Defects}$, not of the form $\alpha \times \text{Start_Date}^\beta + C = \text{Defects}$. Pearson’s coefficient of determination and canonical correlation are optimal for measuring their correlation. The measure extracted for analysis is R^2 .

B.2. Robust regression functions

An attribute group assigned the NUMBER domain class may correlate to an attribute group assigned an INTERVAL domain class in a non-straight-line fashion. We select the Box-Tidwell and Box-Cox functions to measure such a correlation because they provide the best tradeoff between computation speed and accuracy. However, canonical correlation is used in the case of two multi-attribute attribute groups with NUMBER domain classes, because there is no equivalent to the Box-Cox function that handles such a case. R^2 is the measure derived from these functions.

B.3. Ordinal functions

The ordered logit and Spearman's Rho (squared) are the only two functions that exploit order without exploiting distance. The use of Spearman's Rho is discussed in most introductory statistics textbooks (e.g., [28,29]). The semantic equivalence of Spearman's Rho and R is discussed in [54].

B.4. MANOVA and ANOVA

MANOVA and ANOVA are used when one attribute group is assigned the CATEGORICAL domain class and another is assigned the INTERVAL domain class. MANOVA, a generalization of the ANOVA, is used when the attribute group on the interval measurement scale has multiple attributes. The η^2 value of the MANOVA and ANOVA has the same semantic as the regression coefficient R^2 measured by the various regression functions [29].

B.5. Goodman and Kruskal's Lambda

This function transforms the measure of the χ^2 test into a number that is comparable with η^2 and R^2 . Lambda is also used when one attribute group has the ORDINAL domain class, because no function exploits ordering in that situation.

B.6. Logistic regression, point biserial correlation, and Phi coefficient

These functions exploit the pseudo-distance information found in attribute groups in the DICHOTOMOUS domain class to measure correlation.

References

- [1] R. Agrawal, T. Imielinski, A. Swarni, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD Conference on Management of Data, 1993, pp. 207–216.
- [2] P.D. Scott, A.P.M. Coxon, M.H. Hobbs, R.J. Williams, SNOUT: an intelligent assistant for exploratory data analysis, in: Principles of Data Mining and Knowledge Discovery, First European Symposium, PKDD'97, 1997, pp. 189–199.
- [3] W. Hou, Extraction and application of statistical relationships in relational databases, IEEE Transactions on Knowledge and Data Engineering 8 (6) (1996) 939–945.
- [4] A.M. Aladwani, A prototype expert system for selecting a multiple comparison test, International Journal of Computer and Engineering Management 6 (1) (1998), (online) <http://www.journal.au.edu/ijcem/jan98/>.
- [5] C. Böhm, K. Kailing, P. Kröger, A. Zimek, Computing clusters of correlation connected objects, in: ACM SIGMOD Conference on the Management of Data, 2004, pp. 455–466.
- [6] D.J. Hand, Intelligent data analysis: issues and opportunities, in: Proceedings of the Second International Symposium, IDA-97, 1997, pp. 1–14.
- [7] A. Bernstein, F. Provost, Intelligent assistance for the data mining process: an ontology-based approach, IEEE Transactions on Knowledge and Data Engineering, forthcoming.

- [8] A. Bernstein, F. Provost, An intelligent assistant for the knowledge discovery process, in: IJCAI Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases, 2001.
- [9] C. Chua, R.H.L. Chiang, E.-P. Lim, A heuristic method for correlating attribute group pairs in data mining, in: International Workshop on Data Warehousing and Data Mining (DWDM'98), 1998, pp. 29–40.
- [10] C. Chua, R.H.L. Chiang, E.-P. Lim, An intelligent middleware for linear correlation discovery, *Decision Support Systems* 32 (4) (2002) 313–326.
- [11] J.F. Hair Jr., R.E. Anderson, R.L. Tatham, W. Black, *Multivariate Data Analysis with Readings*, fifth ed., Prentice-Hall, 1998.
- [12] W.P. Vogt, *Dictionary of Statistics and Methodology: A Nontechnical Guide for the Social Sciences*, SAGE Publications, 1993.
- [13] S. Brin, R. Motwani, C. Silverstein, Beyond market baskets: generalizing association rules to correlations, in: *Proceedings of the 1997 ACM SIGMOD International Conference on Management of Data*, 1997, pp. 265–276.
- [14] R. Srikant, R. Agrawal, Mining quantitative association rules in large relational tables, in: *ACM SIGMOD Conference on Management of Data*, 1996, pp. 1–12.
- [15] T.D. Cook, D.T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings*, Houghton Mifflin Company, 1979.
- [16] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*, second ed., Lawrence Erlbaum Associates, 1988.
- [17] J.W. Tukey, *Exploratory Data Analysis*, Addison-Wesley, 1977.
- [18] C. Glymour, D. Madigau, D. Pregibon, P. Smyth, Statistical themes and lessons for data mining, *Data Mining and Knowledge Discovery* 1 (1) (1997) 11–28.
- [19] P.S.D. Hand, H. Mannila, *Principles of Data Mining*, MIT Press, 2001.
- [20] U. Fayyad, R. Uthurusamy, Data mining and knowledge discovery in databases, *Communications of the ACM* 39 (11) (1996) 24–26.
- [21] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, *AI Magazine* 17 (3) (1996) 37–54.
- [22] A. Feelders, H. Daniels, M. Holsheimer, Methodological and practical aspects of data mining, *Information & Management* 37 (5) (2000) 271–281.
- [23] D.V. Huntsberger, P. Billingsley, *Elements of Statistical Inference*, Allyn and Bacon, 1987.
- [24] G.D. Israel, Determining sample size, Fact Sheet PEOD-6, University of Florida (November 1992). URL {<http://hammock.ifas.ufl.edu/txt/fairs/13817>}.
- [25] R.S. Lehman, *Statistics and Research Design in the Behavioral Sciences*, Wadsworth Publishing Company, 1988.
- [26] U. Sekaran, *Research Methods for Business: A Skills Building Approach*, John Wiley and Sons, 1992.
- [27] J. Neter, W. Wasserman, M.H. Kutrier, *Applied Linear Regression Models*, second ed., Irwin Homewood, 1989.
- [28] R.B. Burns, *Introduction to Research Methods*, third ed., Addison-Wesley, 1997.
- [29] J. Jaccard, M.A. Becker, *Statistics For the Behavioral Sciences*, second ed., Wadsworth Inc., 1990.
- [30] J. Ziman, *An Introduction to Science Studies: The Philosophical and Social Aspects of Science and Technology*, Cambridge University Press, 1990.
- [31] L.E. Toothaker, *Multiple Comparisons for Researchers*, SAGE Publications, 1991.
- [32] L.E. Toothaker, *Multiple Comparison Procedures*, SAGE Publications, 1993.
- [33] F.M. Wolf, *Meta-Analysis: Quantitative Methods for Research Synthesis, Quantitative Applications in the Social Sciences*, SAGE Publications, 1986.
- [34] J. Neter, W. Wasserman, G.A. Whitmore, *Applied Statistics*, Allyn and Bacon Inc., 1979.
- [35] L.C. Lyons, Meta—analysis: methods of accumulating results across research domains, Available at <http://www.monumental.com/solomon/MetaAnalysis.html> (July 1998).
- [36] D. Donoho, E. Ramos, PRIMDATA: data sets for use with PRIM-H, (online) <http://www.stat.cmu.edu/datasets/> (1982).
- [37] E.R. Berndt, Determinants of wages from the 1985 current population survey, in: *The Practice of Econometrics: Classic and Contemporary*, Addison-Wesley, 1991, pp. 193–209 (chapter 5), [online] <http://www.stat.cmu.edu/datasets/>.
- [38] L.H. Cox, M. Johnson, K. Kafadar, Exposition of statistical graphics technology, in: *ASA Proceedings of Statistical Computation Section*, 1982, pp.55–56.

- [39] A. Heston, R. Summers, The penn world table (mark 5): an expanded set of international comparisons, 1950–1988, *Quarterly Journal of Economics* 8 (6) (1991) 327–368.
- [40] T.J. Biblarz, A.E. Raftery, The effects of family disruption on social mobility, *American Sociological Review* 58 (1993) 97–109.
- [41] V. Greaney, T. Kelleghan, *Equality of Opportunity in Irish Schools*, Educational Company, Dublin, 1984.
- [42] March 1995 population survey—classical families, (online) <http://www.stat.ucla.edu/data/fpp>, 1995.
- [43] R. Kohavi, Scaling up the accuracy of naive-bayes classifiers: a decision-tree hybrid, in: *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 202–207.
- [44] S.E. Fienberg, U.E. Makov, A.P. Sanil, A Bayesian approach to data disclosure: Optimal intruder behavior for continuous data, Tech. Rep. 11/94, Carnegie-Mellon University, 1994.
- [45] A. Alesina, R. Baqir, W. Easterly, Public goods and ethnic divisions, *Quarterly Journal of Economics* 114 (4) (1999) 1243–1284.
- [46] T. Beck, R. Levine, N. Loayza, Finance and the sources of growth, *Journal of Financial Economics* 58 (1–2) (2000) 261–300.
- [47] J.B. De Long, L.H. Summers, How strongly do developing countries benefit from equipment investment?, *Journal of Monetary Economics* 32 (3) (1993) 395–415.
- [48] StatLib, (online) <http://www.stat.cmu.edu/> (1998).
- [49] C. Blake, E. Keogh, C.J. Merz, UCI repository of machine learning databases, (online) <http://www.ics.uci.edu/~mllearn/MLRepository.html>, 1998.
- [50] Economic growth research: published articles and data sets arranged by author, (online) <http://www.world-bank.org/research/growth/paauthor.htm>, 2001.
- [51] L.A. Hayduk, *Structural Equation Modeling with LISREL*, John Hopkins University Press, 1987.
- [52] H. Xiong, S. Shekhar, P.-N. Tan, V. Kumar, Exploiting a support-based upper bound of Pearson’s correlation coefficient for efficiently identifying strong correlation pairs, in: *Proceedings of the Knowledge Discovery in Databases Conference*, 2004, pp. 334–343.
- [53] J. Chilson, R. Ng, A. Wagner, R. Zamar, Parallel computation of high dimensional robust correlation and covariance matrices, in: *Proceedings of the Knowledge Discovery in Databases Conference*, 2004, pp. 533–538.
- [54] M. Kendall, J.D. Gibbons, *Rank Correlation Methods*, fifth ed., Oxford University Press, Oxford, UK, 1990.



Roger Chiang is an Associate Professor of Information Systems at the College of Business, University of Cincinnati. He received his BS degree in Management Science from National Chiao Tung University, Taiwan, MS degrees in Computer Science from Michigan State University and in Business Administration from the University of Rochester, and the PhD degree in Computers and Information Systems from the University of Rochester. His research interests are in data and knowledge management and intelligent systems, particularly in database reverse engineering, database integration, data mining, common sense reasoning and learning, and semantic information retrieval of Web data. He is currently on the editorial board of the *Journal of AIS*, *Journal of Database Management* and *International Journal of Intelligent Systems in Accounting, Finance and Management*. He was the Program Co-Chair of 22nd International Conference on Information Systems, Research in Progress Track, 2001, and ACM International Workshop on Web Information and Data Management in 2001, 2002 and 2003. His research has been published in a number of international journals including *ACM Transactions on Database Systems*, *Data Base*, *Data and Knowledge Engineering*, *Decision Support Systems*, *Journal of Database Administration* and *Very Large Data Base Journal*.



Cecil Eng Huang Chua is an assistant professor at Nanyang Technological University. He received a PhD in Information Systems from Georgia State University, a Masters of Business by Research from Nanyang Technological University and both a Bachelor of Business Administration in Computer Information Systems and Economics and a Masters Certificate in Telecommunications Management from the University of Miami. His research interests include (1) reconstructing the relationship between database and software engineering theories, (2) understanding the role of the IS researcher, and (3) misappropriation of technology. Cecil has several publications in such journals as *Data and Knowledge Engineering*, *Decision Support Systems*, *Journal of the AIS* and the *VLDB Journal*. Cecil maintains the IS Bibliographic Repository, a store of bibliographic information on IS journals.



Ee-Peng Lim is an Associate Professor with the School of Computer Engineering, Nanyang Technological University, Singapore. He obtained his PhD from the University of Minnesota, Minneapolis in 1994. Upon graduation, he started his academic career at the Nanyang Technological University (NTU). In 1997, he established the Centre for Advanced Information Systems and was appointed the Centre Director. He was later appointed a visiting professor at the Chinese University of Hong Kong from December 2001 to June 2003. Upon his return to NTU, he started heading the Division of Information Systems within the School of Computer Engineering. He has published more than 120 referred journal and conference articles in the area of web warehousing, digital libraries and database integration. He is currently an Associate Editor of the *ACM Transactions on Information Systems (TOIS)*. He is also a member of the Editorial Review Board of the *Journal of Database Management (JDM)*. At present, he is the Program Co-Chair of the 2004 Joint Conference on Digital Libraries (JCDL2004) and also the Program Co-Chair of the Sixth International Conference on Asian Digital Libraries (ICADL 2003).