A Continuous Markov-Chain Model of Data Quality Transition: Application in Insurance-Claim Handling

Yuval Zak and Adir Even(✉)

The Department of Industrial Engineering and Management,
Ben-Gurion University of the Negev,
P.O.B. 653, 8410501, Beer-Sheva, Israel

Abstract. Data quality (DQ) might degrade over time, due to changes in real-world entities or behaviors that are not reflected correctly in datasets that describe them. This study presents a continuous-time Markov-Chain model that reflects DQ as a dynamic process. The model may help assessing and predicting accuracy degradation over time. Taking into account cost-benefit tradeoffs, it can also be used to recommend an economically-optimal point in time at which data values should be evaluated and possibly reacquired. The model addresses data-acquisition scenarios that reflect real-world processes with a finite number of states, each described by certain data-attribute values. It takes into account state-transition probabilities, the distribution of time spent in each state, the damage associated with incorrect data that fails to reflect the real-world state, and the cost of data reacquisition. Given current state and the time passed since the last transition, the model estimates the expected damage of a data record and recommends whether or not to correct it, by comparing the potential benefits of correction (elimination of potential damage), versus reacquisition cost.

Following common design science research guidelines, the applicability and the potential contribution of the model is demonstrated with a real-world dataset that reflects a process of handling insurance claims. Insurants' status must be kept up-to-date, to avoid potential monetary damages; however, contacting an insurant for status update is costly and time consuming. Currently the contact decision is guided by some heuristics that are based on employees' experience. The evaluation shows that applying the model has major cost-saving potential, compared to the current state.

Keywords: Data Quality · Accuracy · Continuous-Time Markov Chain · Design Science Research

1 Introduction

Organizations rely on data resources for supporting operations and decision making. As highlighted by a plethora of studies, degradation in data quality (DQ) can be associated with business-process deficiencies, flawed decision and major monetary damages. With the rapid growth in the magnitude of data resources, the task of maintaining high DQ level is becoming increasingly complex and costly, particularly when the detection and the correction of DQ defects require some manual intervention.

B. Donnellan et al. (Eds.): DESRIST 2015, LNCS 9073, pp. 199–214, 2015.
DOI: 10.1007/978-3-319-18714-3_13

DQ management is therefore in a growing need for tools and techniques that can aid and expedite detection and correction in scenarios where the task cannot be fully automated – e.g., by alerting on data items that are likely to be erroneous, predicting possible quality degradation, and improve the cost-effectiveness of manual interventions. The model developed in this study aims at making contribution to that end.

This study addresses scenarios in which data was acquired correctly, but the real-world entity described change over time. If the data is not updated accordingly, it may no longer reflect the real-world state, and becomes inaccurate. For example, if we fail to update a person's data for a while, some attribute might become inaccurate – e.g., the person may have changed address, marital status, or education level. Handling inaccuracies introduce inherent cost-benefit tradeoffs. On the one hand, relying on inaccurate data might lead to fault decisions, possibly associated with some monetary damage. On the other hand, not all accuracy defects can be handled automatically, and manual detection and correction is expensive and resource-demanding. Do the benefits from DQ improvement justify the associated costs? If yes, what is the optimal point in time at which data values should be evaluated and possibly reacquired?

The model developed in this study reflects data-values transitions as a dynamic process. Taking the continuous-time Markov chain approach, the model assumes a finite set of states, each reflecting a possible attribute value. The model also considers the damage caused by inaccuracies – i.e., cases where the data state does not meet the real-world value. As shown later – such formulation can help answer important DQ management questions: a) What is the likelihood that a certain existing data value is inaccurate? b) From the point of acquisition (or, reacquisition) – how long will it take a certain data item to become inaccurate? c) What is the economically-optimal point of time for auditing and possibly correcting a certain data item?

To demonstrate applicability and potential contribution, the model is evaluated with a real-world dataset that reflects insurance-claims handling. Much of the handling is done via phone calls, during which an employee must update the insurant's status. Insurants often neglect to report status updates; hence, the dataset is subject to inaccuracies that translate to major losses for the firm. Contacting all insurants regularly is infeasible, due to time and cost constraints, and currently contact-initiation decisions are guided by heuristics based on employees' experience. The evaluation shows that call-initiation could have substantially improved by applying the model.

In the remainder of this paper, we first review studies that influence our thinking and development. The model formulation is described next, followed by evaluation with real-world data, and discussion of the results. To conclude we summarize the study and its key contributions and highlights possible directions for future research.

2 Background

Data is often subject to quality defects – missing records or values, mismatches between values and real-world entities, outdated values that no longer reflect current behavior, and others. With the broad recognition of data as a critical resource, data quality (DQ) defects and their hazardous effect attract growing attention. Poor DQ may harm operational processes, decision-making activities, and cooperation within and between organizations (Batini, Cappiello et al. 2009). The task of DQ management may involve different perspectives: technical solution, functional requirements, management responsibility, organizational culture, economics, and others (Madnick et al., 2009). This study focuses on DQ assessment - a key DQ management activity (Wang, 1998; Pipino et al., 2002). Measuring DQ and sharing the results can raise awareness to DQ defects, prevent flawed decisions, and help reducing the magnitude of errors and the time spent on validation (Chengalur-Smith, Ballou et al. 1999; Cai and Shankaranarayanan; 2006). A plethora of studies addressed DQ assessment from many different perspectives. Here, we wish to highlight some key questions and insights that rise from a review of some previous works, and influence our study.

2.1 Orthogonal or Dependent DQ Dimensions?

DQ research broadly adopts the notion of DQ dimensions – the claim that DQ should not be assessed as a single "overarching" concept, but rather as a set of perspectives, or dimensions, each reflecting a different type of DQ defects or hazards (Pipino et al., 2002; Even and Shankaranarayanan, 2007) – e.g., Completeness, Accuracy, Currency, Timeliness, and Validity. The common measurement approach, along these dimensions, is a 0-1 ratio that reflects a proportion of non-defected items (1 – perfect DQ, no defects), and can be assessed at different levels - records, specific attributes, or entire datasets (Pipino et al., 2002; Even and Shankaranarayanan, 2007).

A first question that we raise is – should DQ dimensions be treated as orthogonal or dependent? So far, DQ dimensions were more commonly treated as orthogonal and assessed independently. This approach is apparent in works that discuss a specific dimension (e.g., Even et al., 2010; Fisher et al., 2009; Heinrich and Klier, 2011; Wechsler et al., 2013), or multiple dimensions, each measured independently (Pipino et al., 2002; Even and Shankaranarayanan, 2007). Some studies, however, look at possible mutual effect between DQ dimensions – how changes in one are reflected in others. Ballou and Pazer (1995; 2003) look into accuracy-timeliness and completeness-consistency tradeoffs. Parssian et al. (2004) analyze the evolvement of DQ defects along a multi-stage process, showing that defects of a certain type may evolve into defects of other types at later stages.

This study looks into the mutuality between currency and accuracy – the former reflects the extent to which data is up-to-date, while the latter reflects the extent to which the data is free of errors. It shows that as data becomes less current it is also likely to become less accurate. A similar proposition was made by Wechsler et al. (2012), who developed a model that highlights possible mechanism behind that mutual effect and demonstrated it with census data. This work will be discussed some more later, as it influenced the conceptualization the model development in this study.

2.2 DQ as a Static Snapshot or as a Dynamic Process?

DQ measurements serve as input for important DQ management tasks – analysis of current state, communicating DQ status to end-users, and directing improvement efforts (Wang, 1998). *A second question that we raise is – should assessment take a static ("Snapshot") view, or rather a dynamic view of DQ as an evolving process.*

Many works reflect a static view – assessment based on a "snapshot" of data, taken at a certain point of time (e.g., Ballou et al., 1995; Chengalur-Smith et al., 1999; Even et al., 2010). Some works introduce a probabilistic approach into their measurement schema, acknowledging the fact that the data sample available for evaluation does not necessarily provide a comprehensive and recent enough picture of the real-world state (e.g., Fisher et al., 2009, Heinrich, et al., 2009/2011). Regardless the probabilistic approach, this body of works still offers a static view, and provider measures that reflect the current DQ state.

A dynamic view is reflected to an extent in Pipino et al. (2002) – their software utility permits tracking progression of "snapshot" measurements over time. Even et al. (2010) show that DQ deteriorates over time, to a point where outdated data might become useless and no-longer worth fixing. Wechsler et al. (2012) model transitions between data values as a multi-stage dynamic process that explains DQ deterioration. An important motivation behind dynamic modeling is the possibility to turn it toward prediction of future DQ degradation. If predictions are reasonably reliable, managers can prepare for possible DQ hazard, act proactively, and take preventive measures.

2.3 Impartial or Value-Driven DQ Assessment?

A number of studies have highlighted economic aspects of DQ. A possible perspective for observing economic DQ issues is the value of information, as high-quality data is positively associated with higher value or utility (Haug et al., 2011, Even and Shankaranarayanan 2007). DQ defects might degrade the potential value, and cause monetary losses – e.g., by resulting-in sub-optimal decisions (Heinrich, Klier et al. 2009, Even et al. 2010). The other possible economic perspective is the cost associated with DQ improvement – manual handling of DQ defects typically require major time resources (Wechsler and Even, 2012), while automation require investment in IT resources (Cappiello et al. 2003, Eppler and Helfert, 2004).

Our third question – should the goal of DQ assessment be error-free data, or maximizing value and economic benefits? Even and Shankaranarayanan (2007) link this differentiation to impartial versus contextual DQ measurement. The former reflects stand-alone assessment of data and DQ defects, regardless of how data is used. The latter reflects the impact of DQ defects within a specific context of use. Their contextual assessment applies the concept of utility – a measure for the value stems for data usage that may vary, depending on the usage contexts. Impartial measurement is more common in earlier DQ works (e.g., Ballou and Pazer, 1995/2003, Chengalur-Smith et al., 1999; Parssian et al., 2004), while some more recent works look into linking DQ assessment to data utilization with the associated benefits (e.g., Even et al., 2010; Heinrich et al., 2009; Wechsler et al., 2013). This study links DQ assessment with the utility damage of inaccuracies and the cost of correction, toward economically-optimal prioritization of DQ improvement efforts.

3 Model Development

The model developed in this section addresses data management scenarios that adhere to the following characteristics and assumptions:

- A dataset, in which each record reflects a single instance (e.g., a list of customers).
- A target attribute, with value that reflects the real-world state of the associated instance (e.g., the customer's status). The model assumes a finite number of real-world states, each associated with a corresponding data value. Hence, the value domain of the target attribute is a discrete and finite set of possible values.
- When a dataset record is added or updated, the target attribute reflects correctly the real-world state. However, the real-world state may change, and if the target attribute is not updated accordingly, it no longer reflects the real-world state accurately.
- The target attribute is assumed to have significant business importance, with a certain cost or penalty in case of inaccuracies; hence, the motivation for maintaining target-attribute values as accurate as possible.
- Besides the target attribute, a record contains a number of additional attributes (e.g., the customer's gender, date of birth, or region of residence). Some of those attributes may have some association with the target attribute, and may help predicting transitions in the real-world state to an extent.

A modeling approach that may fit such scenarios is the Markov Chain (MC) model (Ross, 1996). The basic MC form considers a stochastic process of transitioning over time between a finite number of possible values $\{x_i\}_{i=1..N}$. Time is modeled as a discrete variable (t = 0, 1, 2, ..), where steps in [t] are associated with equal time interval. The transition probability P_{ij} reflect the likelihood of transitioning from value x_i to value x_j within a single time interval. The MC assumes "memory-less" transitions – i.e., the transition probability depends only on the current value, and not on previous values, and does not change over time. The collection of transition probabilities forms the transition matrix, where $\sum_{j=1..J} P_{i,j} = 1$ for each [i]:

$$P = \begin{bmatrix} P_{11} & \cdots & P_{1n} \\ \vdots & \ddots & \vdots \\ P_{n1} & \cdots & P_{nn} \end{bmatrix} \quad (1)$$

P is assumed to be stationary; hence, $P(t)$, the t-steps transition matrix (i.e., the set of probabilities that a value will change from x_i to x_j after t periods) is the t-power of the transition matrix: $P(t) = P^t$.

A model for DQ assessment, based on the basic MC form, was introduced in (Wechsler and Even, 2012). The proposed model fits the characteristics and the assumption of the scenario described above. If a certain target-attribute value x_i was recorded at time $t=0$, it can be shown that its expected accuracy level (the likelihood to remain accurate) at a later time t is given by $A_i(t) = P_{ii}(t)$. The modeling approach proposed by that study had major influence on the approach applied in this study. However, that modeling approach poses a few major limitations, which are addressed by this study: fixed-length discrete time periods, possible dependencies between attributes, and the need to consider possible cost-benefit tradeoffs.

Notably, extended MC forms offer refined treatment of time (Ross, 1996). The *Continuous-Time Markov Chain* extends the MC to a continuous stochastic process. In the continuous-time MC the time spent in state x_i has the "memoryless" property as well. Let τ_i denote the time spent in state x_i before transitioning, then $P\{\tau_i > s + t | \tau_i > s\} =$

$P\{\tau_i > t\}$. The random variable τ_i must therefore be exponentially distributed. The transition probability from x_i to state x_j depends of the transition time: $P_{ij}(t) = P\{X(t+s) = j|X(s) = i\}$. This MC extension is used in the development of our model, which is described next.

3.1 Baseline Formulation: Optimal Data Reacquisition Time

Next, we present an analytical formulation, aimed at answering the question: given a record with a target-attribute value of x_i, what would be the optimal time for reacquisition of that record? The formulation, which considers the following factors, is first stated at a high-level, and further extended later:

- The time, denoted by (t), passed since the most recent data acquisition,
- The real-world property, reflected by the target attribute has *N* possible states. Accordingly, the target attribute has one among *N* possible values $\{xi\}_{i=1\ldots N}$
- The data state, as reflected by a current target-attribute value x_i, (denoted by index [i]) vs. the real-world state, which should have resulted a target-attribute value x_j (denoted by [j]). The record is said to be accurate if i=j, and inaccurate otherwise.
- Inaccuracy may result in some monetary damage, which may change over time. The damage function $d_{ij}(t)$ reflects the damage that can be attributed to a record currently at state [i], which should have been in state [j]. All $\{d_{ij}(t)\}$ are assumed to be non-negative, monotonic and non-decreasing with (t) (i.e., can be a constant). We also assume no damage when a record is accurate (i.e., $d_{ii}(t) = 0$). Since the assumption if that at the time of acquisition the data is accurate, $d_{ij}(0) = 0$.
- The cumulative damage function D_i reflects the accumulation of damage functions, weighted by the probability or transitions. $P_{ij}(t)$ reflects the probability that a record with state [i] at t=0 has transitioned to state [j] at time t. The cumulative function can therefore be expressed as $D_i(t) = \Sigma_{j=1..N}\left(P_{ij}(t) \cdot d_{ij}(t)\right)$. Since $D_i(t)$ is a linear combination of $\{d_{ij}(t)\}$ with non-negative weights, it is also non-negative, monotonic/non-decreasing with (t), and $D_i(0) = 0$.
- Reacquisition cost $C_i(t)$ depends on current state [i] and on (t). As with the damage function, we assume that $C_i(t)$ is positive, monotonic, and non-decreasing with (t).

The optimal point of time for data reacquisition is (t) that solved the following:

$$C_i(t) = D_i(t) = \sum_{j=1\ldots N}\left(P_{ij}(t) \cdot d_{ij}(t)\right) \tag{2}$$

In other words: data reacquisition should be performed at the first point of time where the potential damage is higher than the reacquisition cost.

- Since $C_i(t)$ and $D_i(t)$ are both monotonic and non-decreasing, there is at the most one optimal point of time t_{opt} that solves the equation (Fig. 1a).
- At the time of acquisition (t=0), the cost is positive; hence, greater than the damage: $C_i(0) > D_i(0) = 0$. If for all (t) $C_i(t) > D_i(t)$ (Fig. 1b), no reacquisition will occur. This is particularly true if the record is at a state with no transitioning out (a "sink") – i.e., $P_{ij}(t)$ is 1 for i=j, 0 otherwise, and $D_i(t) = 0$ for all (t).

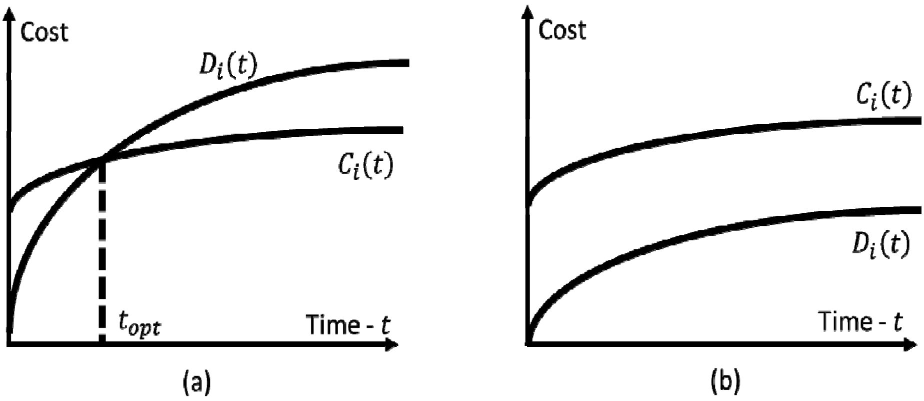

Fig. 1. Inaccuracy Damage versus Correction Cost

3.2 State Transitions as a Continuous-Time Markov Chain

Assuming that the transitions between states follow a Continuous-Time Markov-Chain (CTMC), the formulation in Eq. 2 can be further extended:

- The probability of transition $P_{ij}(t)$ from state [*i*] to state [*j*], as a function of time (*t*), is seen as assembled from two separate components: the probability of transition, and the time between transitions.
- Assuming a CTMC, the probability of transition from state [*i*] to state [*j*] is a constant P_{ij} - a component in the transition probability matrix *P*, described earlier in Eq. 1. This probability does not dependent on the previous states in which the records resided, other than the current state [*i*].
- Assuming a CTMS, the time between transitions is exponentially distributed, dependent on the current state [i], and the next state [j]. It is defined as ~exp $(\lambda_{ij}\alpha_{ij})$, where λ_{ij} is the transition rate from state [*i*] to state [*j*], and α_{ij} is an auxiliary parameter.
- Assuming an exponential distribution, the probability of a record to transition out of state [*i*] by the time (*t*), given target state [*j*], is $P(t|i \rightarrow j) = 1 - e^{-\lambda_{ij}\alpha_{ij}}$.
- The damage function $d_{ij}(t)$ is based on assessment of potential damages. The formulation has to be defined over the $[0, \infty]$ range, and adheres to the conditions defined earlier: non-negative, monotonic, and non-decreasing with (*t*)
- Since the damage function $d_{ij}(t)$ is zero at *t = 0* ($d_{ij}(0) = 0$), the expected potential damage (i.e. the accumulative damage) at time (*t*) is defined as $E_D[t] = \int_0^t D_i(x)dx = D_i(t) - D_i(0) = D_i(t) - 0 = D_i(t)$

Based on these assumptions, the formulation presented in Eq. 2 of the optimal point for the time for data reacquisition, can be extended to:

$$C_i(t) = D_i(t) = \sum_{j=1...N}\left(P_{ij} \cdot \left(1 - e^{-\lambda_{ij}\alpha_{ij}}\right) \cdot d_{ij}(t)\right) \tag{3}$$

3.3 Context Variables

The scope of assessment and correction is the target attribute but it would be important to mention that the data record contains additional attributes. Obviously, such attributes may have important business value too and their quality must be managed as well. Here, for the purpose of the model development, we see them as *context attributes* – they may have some association with the target attribute, describes certain relevant conditions under which it transitioned from one state to another, and possibly have some influence on the different components of the model formulation: the damage functions, the transition probabilities, and/or the correction costs. We now extend the formulation to reflect the possible impact of the context attributes:

- We assume a single context attribute with a value domain that contains a discrete set of L possible values or states $\{x_l\}_{l=1...L}$, indexed by [l]. Each state reflects a certain context that classifies the records into meaningful groups (e.g., customer segments, system of data origin, etc.).
  - A context variable defined over a continuous range (e.g., "annual salary"), can be transformed into a discrete set of ranges ("bins") that reflect meaningful business classification (e.g., "salary range", of "high", "medium", or "low").
  - Multiple attributes can be transformed to a single context attribute, in which each state reflects a combination of value. For example, a combination of {"marital status" and "salary range"} with possible value combinations of ("single", "low"), ("single", "medium"), ("married", "low"), etc.
- The assumption is that the context value of a record is set when the record is first acquired, and does not change over time. As discussed in the concluding section, later extensions to this work should look into modeling possible transitions in [l].
- The specific context value [l] may affect all the model components; hence, their annotation should be extended accordingly:
  - The correction cost function $C_{il}(t)$
  - The damage function $d_{ijl}(t)$ and the cumulative damage function $D_{il}(t)$.
  - The transition probability matrix P_l and its cells $\{P_{ijl}\}$.
  - The exponential distribution parameters: ~exp $(\lambda_{ijl}\alpha_{ijl})$.

Accordingly, the formulation in Eq. 3 should be extended to:

$$C_{il}(t) = D_{il}(t) = \sum_{j=1...N}\left(P_{ijl} \cdot \left(1 - e^{-\lambda_{ijl}\alpha_{ijl}}\right) \cdot d_{ijl}(t)\right) \tag{4}$$

The implication is that, given a context variable with a set of possible states, the model has to be evaluated separately for each state. This implies a need to establish L models, one for each context state, and for each record apply the model that matches the context group to which it belongs.

4 Empirical Evaluation

This study contributes a novel model for predicting economically-optimal data reacquisition cost. This contribution aligns with the design-science research (DSR) paradigm, which targets the creation of new artifacts (such as models) toward improving IS implementation. The success of DSR outcome is judged by its quality, contribution, and the impact of the developed artifacts (Hevner at al., 2004). The work described so far can be linked to the DSR steps defined in (Peffer et al., 2007):

1. *Identify Problem and Motivate:* data inaccuracies may cause substantial damages, and the cost of fixing them may turn out to be high, hence the need for solutions that may help predicting inaccuracies, assess the potential damage, and prioritize improvement efforts accordingly.
2. *Define Objectives and Solutions:* the proposed solution is an analytical model, based on the Markov-Chain approach, which helps predicting accuracy degradation, and help assessing the cost-benefit tradeoffs associated with fixing it.
3. *Design and Development:* the development of the proposed model was described in the previous section.

In this section we proceed to the next steps: demonstration of the model and evaluating its performance within a suitable real-world data management scenario.

4.1 Evaluation Setup – The Firm and the Business Process

Our evaluation site is a privately-owned service provider (referred to as the FIRM) that works in collaboration with leading Health Maintenance Organizations (HMO's) and handles insurance claim for customers who suffered work accidents (magnitude of 10,000's claims, annually). A person who suffers an accident is entitled for some benefits (e.g., monthly stipend for the recovery period, coverage of medical expenses, and help in transportation) from the National Insurance Organization (NIO). The process of applying for those benefits is long and complex, and required submission of applications, medical records, and specialist assessments. It is in the interest of the HMO's that a customer fills-in the application, otherwise medical expenses that could have been covered by the NIO, will have to be charged to the HMO's. The HMO's hire the FIRM to accompany the customer, assist them with the claim-application process and make sure that the required documentation is delivered. Customers are not charged for this service, and the FIRM is getting reimbursed for claims that ended-up being filed.

The process is mostly remote – i.e., almost no face-to-face meetings are required, and most of the status tracking is done via phone call with serviced representatives. The claims are filed either by FIRM representatives or by the customer. Customers are supposed to report FIRM representative on any progress. However, in practice, they often neglect to do so; hence, their data record often does not reflect correctly their actual status. Discrepancies as such might turn out to be costly – for example, if the customer has received the forms, but failed to complete and sign them, the processes might be substantially delayed, and the FIRM will not get reimbursed. To avoid possible discrepancies, FIRM representatives call customers that are still in the process eventually, and verify their status. Making such a call costs the service-representative time; hence, cannot be performed too often. To avoid too-high cost, the representatives call only a subset of the customers each month, where the choice is based on their current state, and other "heuristics" that have evolved in the firm over the years. Given this current situation of severe damages due to data inaccuracies, and high data reacquisition cost, the FIRM is currently looking into a solution that will help turning the customer calls into a more cost-effective process.

With the help of FIRM managers, the claim-handling was modeled as an 8-stage process with possible transitions among them.

1. *Customer data received***:** data was received from the HMO, no contact with the customer was made yet.
2. *Customer is waiting for the forms*: first contact with the customer was made, and the customer is waiting for the forms.
3. *Customer is filling the forms*: customer has received the forms and needs to fill them and get his employer to sign.
4. *Customer had signed the forms*: customer had filled-in and signed the forms, and needs to deliver it back to the company.
5. *Claim was filed*: the claim was submitted to the NIO, and pending for processing and approval.
6. *Claim was filed independently*: the customer had chosen to file the claim independently, with only partial help of the company.
7. *Process is irrelevant*: the customer is either unreachable, had already filed in a claim, or is interested in filing a claim at all.
8. *Claim approved*: the claim processing by the NIO has been completed, and the application was approved.

From the FIRM's stand point – the hazardous states, with some potential damage, are 6 and 7. In all the stages the customer is associated with some cost (the time spent on calls so far), but the FIRM will see no revenue.

4.2 Data Collection and Preparation

The evaluation included a dataset with 14,209 customer records. The records were anonymized – Id's were converted to sequential numbers, and any detail that could have identified the customer was removed. The current process step (a value between 1 and 8) was defined as the target attribute. As context attribute we chose an attribute that reflected three forms (L=3) of how the contact with the customer is handled: phone calls (l=1, 7,513 records), field representative (l=2, 5,330 records), or a combination of both (l=3, 1,366 records). FIRM managers suggested that the contact form has important implications for the process, and significant impact on costs and potential damage; hence three models were developed, one for each context value.

Currently, the FIRM's customer database does not keep track of the changes, and does not record the exact date and time in which the status was updated. To track changes in status, we sampled the dataset periodically, and compared customer status and the beginning and at the end of each period. Overall, we sampled in periods, each reflecting a slightly different number of days (32, 31, 34, 30, 32, and 29). Transition matrices $\{P\}$ for all models were calculated for each period – we have verified and the transition probabilities were indeed similar between periods; hence, it was reasonable to assume that the transition matrices stay stationary over time. This sampling schema introduced some issues that had to be addressed in the evaluation:

- The dataset state the last date of contacting the customer. It is therefore possible to know whether or not the customer was contacted during the month, and if yes – whether or not the call resulted in a change in state. However, as only the last call is recorded, it is impossible to tell whether or not within a single period a customer received a few calls, and the status updated more than once. The evaluation therefor considered only the last transition within a period, if more than one occurred. Since the model assumes memoryless transitions, this approximation did not bias significantly the model outcomes.
- If a customer was not contacted during the period – it is impossible to tell whether or not the real-world state did change. In that case, a possible transition in the real world state had to be approximated – based on actual transitions of customers at similar states, who were contacted via the same contact form.

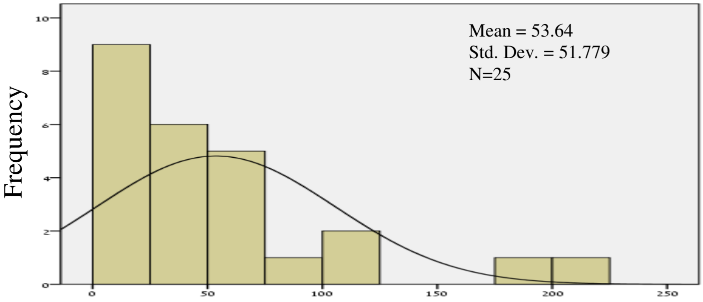

Fig. 2. Transition Histogram Example: From State 1 to State 8 under Contact From 1

The assessment of probability parameters was conducted for all 6 periods, where the estimation was conducted for customers who performed transitions. The transition probability matrices reflected, in general, the expected business process – for example, state 8 that was expected to be stationary, indeed did not have transitions out. However, the transitions did not always conform to the assumption of exponentially distributed transition time. Out of 132 transition rates, only 64 were shown to be exponentially distributed (using the K-S goodness-of-fit test, with significance level of 0.05). Some transition time distributions did not pass the test, but still appeared to have nearly-exponential characteristics (For example – the distribution shown in Fig. 2). Despite the misfit of some distributions to the model assumption, we chose to proceed with the model, in hope that when applied it can still yield better prediction results than current performance, in terms of cost-saving.

4.3 Evaluation Results

The model can help predicting the optimal time in which data should be reacquired. The potential damage of data inaccuracy is zero at *t*=0 and may grow over time. The evaluation was conducted, for each period, along the following steps:

- The customers who were evaluated were those with known state at the beginning of the period (i.e., not newly-added), and not stationary (i.e., state other than 8).
- Per customer, the model was used to predict the potential damage at the end of the period. From the FIRM's standpoint, the damage will realizes if the customer actually reached states 6 or 7 – but the data shows a different state, not perceived as hazardous. The potential damage was therefor set as the likelihood that a customer will reach one of the hazardous states at the end of a period, given current state. The records were there sorted by their potential damage – high to low.
- The evaluation compared the performance against the current heuristics-based calls by FIRM representatives. If at a certain period Y calls were made – the evaluation compared Y customers who were actually called to the Y customers that that were ranked as having the highest potential damage according to the model's prediction.

Table 1. Prevention of Potential Damage, Actual vs. Model

Pd.	Recor ds	Potential Damage	Potential Damage Prevented – Actual	Potential Damage Prevented – Model	Improvement Percentage
1	4608	561.01	516.96 (92%)	558.17 (99%)	**8%**
2	3943	555.32	355.81 (64%)	488.44 (88%)	**37%**
3	3385	676.98	267.85 (40%)	531.54 (79%)	**98%**
4	5412	597.92	543.53 (91%)	592.69 (99%)	**9%**
5	4123	592.4	371.88 (63%)	512.20 (86%)	**38%**
6	3315	673.94	285.35 (42%)	511.40 (76%)	**79%**

Table 1 compares the potential damage prevented, which is defined as the potential damage of a customer who were actually called by representatives (or recommended by the model). In all periods, the predictions made by the model could have prevented more potential damage. The margin is explained by the quality of recommendations made by the model – while both methods were evaluated with the same number of customers per period, the model could recommend customers with higher damage potential to be contacted.

Table 2 demonstrates how the suggested model prevents actual damage, by comparing model recommendation to the transition during the evaluated period.

- Damage prevented: customers who were actually called by representatives (or recommended by the model) and ended-up transitioning. When the customer's state is reacquired, if transitioned to states 6 or 7, the damage was considered as damage prevention.
- Damage inflicted: customers who were not called by representatives (or not recommended by the model), and ended-up transitioning to another state.
- Except for period 1, the model increased the damage prevented and reduced the damage inflicted. In some periods major improvements were made.

Table 2. Damage Analysis

Pd.	**Recs.**	**Potential Damage**	***Damage Prevented – Actual***	**Damage Prevented – Model**	***Prevention Increase (%)***	***Damage Inflicted – Actual***	**Damage Inflicted – Model**	***Infliction Decrease (%)***
1	4608	561.01	400.06 (71%)	394.51 (70%)	**-1%**	41.00 (7%)	46.55 (8%)	**14%**
2	3943	555.32	308.24 (51%)	329.29 (54%)	**7%**	216.35 (36%)	195.30 (32%)	**-10%**
3	3385	676.98	198.95 (27%)	311.51 (42%)	**57%**	335.70 (46%)	223.14 (30%)	**-34%**
4	5412	597.92	365.23 (61%)	386.68 (64%)	**6%**	56.23 (9%)	34.78 (6%)	**-38%**
5	4123	592.4	294.80 (45%)	354.48 (54%)	**20%**	239.30 (37%)	179.61 (27%)	**-25%**
6	3315	673.94	243.92 (33%)	326.84 (44%)	**34%**	329.27 (45%)	246.35 (34%)	**-25%**

The context attribute that we chose for evaluation is the form of contact, with 3 possible values. As suggested earlier (Eq. 4), the model parameters were developed for each form separately and the customer subgroups where evaluated each according to the associated model. The evaluation in Table 2 has been repeated, but disregarding the context-value. The results are summarized in Table 3.

The non-context-evalution results, as presented in Table 3, are fairly similar to the evalution that did consider the differences in context. In some periods the damage prevention was higher and the damange inflinction was smaller, but in average the preformacnce was similar, with no statistically-significant difference. Without contextual attribute the average damage prevention was 21.53% and the averatge damage infliction was -22.2%, while when splitting the customers into 3 groups the average numbers were 20.36% and -19.67%, respectively.

Overall, the evaluation results were encouraging. The use of the model was able to provide recommendations of customers with high probability of state transition that need to be contacted, with overall performance that was substantially higher than the current heuristics-based contact method. A key preliminary assumption, made prior to the model evaluation, was that the transition time has exponential distribution. This assumption was supported only partially by the actual data – some distributions confirmed this assumption but some did not. Regardless – the use of the model could provide good results, in terms of damage reduction, despite some mismatched with the assumption of exponential behavior. The further separation to different models, based on the values of the contact form as context variable, did not improve the results but did not harm them either. Notably the choice of context variable was based on a recommendation made by FIRM's managers. A more robust evaluation of context variable is needed, and a better choice could have possibly made a greater impact.

Table 3. Damage Analysis, Disregarding the Impact of Context Attribute.

Pd.	**Records**	**Potential Damage**	***Damage Prevented – Actual***	**Damage Prevented – Model**	***Prevention Increase (%)***	***Damage Inflicted – Actual***	**Damage Inflicted – Model**	***Infliction Decrease (%)***
1	4608	452.49	363.88 (80%)	363.81 (80%)	**0%**	39.79 (9%)	39.86 (9%)	**0%**
2	3943	451.37	281.41 (62%)	284.12 (63%)	**1%**	195.6 (43%)	192.89 (43%)	**-1%**
3	3385	517.28	182.47 (35%)	304.86 (59%)	**67%**	301.87 (58%)	179.48 (35%)	**-41%**
4	5412	523.15	333.86 (63%)	360.53 (69%)	**8%**	59.39 (11%)	32.71 (6%)	**-45%**
5	4123	513.21	269.84 (52%)	306.75 (60%)	**14%**	215.76 (42%)	178.85 (35%)	**-17%**
6	3315	520.55	223.75 (42%)	312.11 (60%)	**39%**	300.54 (58%)	212.18 (41%)	**-29%**

5 Conclusions

With the growing dependency of organization on their data resources, the issue of data quality defects and their potential damage is on the rise. Data quality management is in need for tools and techniques that will aid the associated decisions – which data items should be audited and possibly corrected, what is the optimal timing to do so, and how to do so in a cost-effective manner. This study contributes to that end by offering a model that can help predicting possible degradation in data quality and recommending the optimal time for requisition. The model, based on a continuous-time Markov chain, takes some novel approaches, compared to tools and techniques that were previously introduced in research. The study looks at a possible interplay of two DQ dimensions that are mostly treated independently – accuracy and currency. It observes quality degradation as dynamic process, and builds into the model possible cost-benefit tradeoffs, that can influence economically-optimal choices.

The application and the potential contribution of the model were demonstrated with a large dataset that reflects a dynamic real-world scenario with characteristics that justify, in general, the model formulation and assumptions. The results were encouraging, and the model indeed showed a potential to improve the data acquisition process and reduce damage. Obviously – some more evaluation and adjustments are required, before the model can turn into a tool that can be applied in practice.

While this study makes some contributions, it had some limitations that should be acknowledged, and possibly addressed in future research. The model relies on the assumption that the company reacquires the customers' current state without interfering with the natural course of their process. In practice, contacting a customer for data reacquisition may serve as an opportunity to make some offers and influence the customers' behavior. By that – the act of data acquisition does not only update the record to reflect the real-world state, but can also influence the real-world state and result in some changes to the data. Modeling reacquisition as a decision tree may help capturing this possibility. Another limitation is the underlying assumption of a memory-less transition time with exponential distribution. In this study, this assumption was applied even in cases where the actual transition time did not match an exponential behavior. In future extensions, the model can be further developed to deal with different type of distribution. A third limitation is the assumption that context variable are stationary – i.e., their value is set when the record is first acquired, and does not transition over time. Obviously, this assumption applied only with certain context attributes, but not with others. A future enhancement to the model can consider possible transitions in the values of context attributes, and assess the possible impact of such transitions on the ability to predict the transition time, and optimize the reacquisition decision accordingly. The improvements discussed here are currently under development and evaluation, and we plan to present them in a follow-up study.

References

1. Ballou, D.P., Pazer, H.L.: Modeling completeness versus consistency tradeoffs in information decision contexts. IEEE Trans. Knowledge and Data Eng. 15(1), 240–243 (2003)
2. Ballou, D.P., Pazer, H.L.: Designing information systems to optimize the accuracy-timeliness tradeoff. Information Systems Research 6(1), 51–72 (1995)
3. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. ACM Computing Surveys (CSUR) 41(3), 16 (2009)
4. Cai, Y., Shankaranarayanan, G.: Supporting data quality management in decision-making. Decision Support Systems 42(1), 302–317 (2006)
5. Cappiello, C., Francalanci, C., Pernici, B.: Time-related factors of data quality in multichannel information systems. J. of Management Information Systems 20(3), 71–92 (2003)
6. Chengalur-Smith, I.N., Ballou, D.P., Pazer, H.L.: The impact of data quality information on decision making: An exploratory analysis. IEEE Transactions on Knowledge and Data Engineering 11(6), 853–864 (1999)
7. Eppler, M., Helfert, M.: A classification and analysis of data quality costs. Paper presented at the International Conference on Information Quality (2004)
8. Even, A., Shankaranarayanan, G.: Utility-driven assessment of data quality. ACM SIGMIS Database 38(2), 75–93 (2007)
9. Even, A., Shankaranarayanan, G., Berger, P.D.: Evaluating a model for cost-effective data quality management in a real-world CRM setting. DSS 50(1), 152–163 (2010)
10. Fisher, C.W., Lauria, E.J., Matheus, C.C.: An accuracy metric: Percentages, randomness, and probabilities. Journal of Data and Information Quality (JDIQ) 1(3), 16 (2009)
11. Haug, A., Zachariassen, F., Van Liempd, D.: The costs of poor data quality. Journal of Industrial Engineering and Management 4(2), 168–193 (2011)
12. Heinrich, B., Klier, M., Kaiser, M.: A procedure to develop metrics for currency and its application in CRM. Journal of Data and Information Quality (JDIQ) 1(1), 5 (2009)
13. Heinrich, B., Klier, M.: Assessing data currency—a probabilistic approach. Journal of Information Science 37(1), 86–100 (2011)
14. Hevner, A.R., March, S.T., Park, J., Ram, S.: Design Science in Information Systems Research. MIS Quarterly 28(1), 75–105 (2004)

15. Madnick, S.E., Wang, R.Y., Lee, Y.W., Zhu, H.: Overview and framework for data and information quality research. J. of Data and Information Quality (JDIQ) 1(1), 2 (2009)
16. Parssian, A., Sarkar, S., Jacob, V.S.: Assessing data quality for information products: Impact of selection, projection, and Cartesian product. Management Science 50(7), 967–982 (2004)
17. Peffers, K., Tuunanen, T., Rothenberger, M., Chatterjee, S.: A Design Science Research Methodology for Information Systems Research. Journal of Management Information Systems 24(3), 45–77 (2007)
18. Pipino, L.L., Lee, Y.W., Wang, R.Y.: Data quality assessment. Communications of the ACM 45(4), 211–218 (2002)
19. Ross, S.M.: Stochastic processes, 2nd edn. Wiley, USA (1996)
20. Wang, R.Y.: A product perspective on total data quality management. Communications of the ACM 41(2), 58–65 (1998)
21. Wechsler, A., Even, A.: Assessing accuracy degradation over time with A Markov-chain model. In: The 17th Intl. Conference on Information Quality (ICIQ), Paris (2012)
22. Wechsler, A., Even, A., Weiss-Meilik, A.: A Model for Setting Optimal Data-Acquisition Policy and its Application with Clinical Data. In: The Intl. Conf. on Information System (ICIS), Milan, Italy (2013)