Embedding Forecast Operators in Databases

Francesco Parisi[1], Amy Sliva[2,3], and V.S. Subrahmanian[2]

[1] Università della Calabria, Via Bucci–87036 Rende (CS), Italy
fparisi@deis.unical.it
[2] University of Maryland College Park, College Park MD 20742, USA
vs@umiacs.umd.edu
[3] Northeastern University, Boston MA 02115, USA
asliva@ccs.neu.edu

Abstract. Though forecasting methods are used in numerous fields, we have seen no work on providing a general theoretical framework to build forecast operators into temporal databases. In this paper, we first develop a formal definition of a forecast operator as a function that satisfies a suite of forecast axioms. Based on this definition, we propose three families of forecast operators called *deterministic*, *probabilistic*, and *possible worlds* forecast operators. Additional properties of coherence, monotonicity, and fact preservation are identified that these operators may satisfy (but are not required to). We show how deterministic forecast operators can always be encoded as probabilistic forecast operators, and how both deterministic and probabilistic forecast operators can be expressed as possible worlds forecast operators. Issues related to the complexity of these operators are studied, showing the relative computational tradeoffs of these types of forecast operators. Finally, we explore the integration of forecast operators with standard relational operators in temporal databases and propose several policies for answering forecast queries.

1 Introduction

Though time series analysis methods have been studied extensively over the years in many contexts [3], there has only recently been work that merges classical forecasting with standard operations in temporal databases [1,5,6]. Given the widespread use of temporal data, there are numerous applications that require such capabilities, allowing for the consistent use and application of time series forecasts in databases. A university might want to forecast research grant income (or expenditures) in the future by examining a database of research projects. A stock market firm might want to include support for various kinds of specialized forecasting algorithms that predict the values of mutual fund portfolios or a single stock over time. A government might want to forecast the number of electricity connections or other development indicators in their country over time. Such forecasts might not just be made about the future, but also used to fill in gaps about the past. For instance, using data about the number of electricity connections in Ecuador from 1990–2000 and 2002–2007, officials may want to interpolate the number of connections there might have been in 2001.

This paper is not about how to make such forecasts. Currently, in all forecasting applications, the model building and forecasting is performed outside of the database

S. Benferhat and J. Grant (Eds.): SUM 2011, LNAI 6929, pp. 373–386, 2011.

system itself, rather than as a smoothly integrated process. The implementation of these forecasting models are often ad hoc in nature, and general relationships between different forecasting tasks and domains are not exploited to their full potential. Yet, the broad demand for forecasting and predictive analyses creates a need for a robust theoretical framework that can incorporate forecasting directly into temporal databases.

The field of forecasting is extensive and widely studied, with an array of general techniques [3] as well as specialized forecast models for a variety of domains, such as finance [15], epidemiology [9], politics [2,10,11,14], and even product liability claims [13]. All these methods are very different from one another, and even within a restricted domain such as the stock market, there are hundreds of forecasting models available, each with varying strengths and weaknesses. In spite of these variations, we can identify general properties of forecasting that will facilitate integration of these methods into query languages, making them available for managing and analyzing temporal databases.

In this paper, the question *"what should count as a forecast operator*" is answered by first providing a set of axioms that a forecast operator must satisfy. We assume that forecast operators apply to temporal relational databases—the main reason for this assumption is that in today's world, most (though certainly not all) temporal data is in fact stored in such databases. Subsequently, we define three classes of forecast operators—deterministic forecast (DF) operators, probabilistic forecast (PF) operators, and possible worlds forecast (PWF) operators. We show that DF operators are a special case of PF operators which in turn are a special case of PWF operators. Certain classical forecasting methods such as linear regression, polynomial regression, and logistic regression methods are all demonstrated to be special cases of this framework. Some new operators for forecasting are also developed, along with results characterizing the complexity of applying certain forecast operators. This generalized understanding of the properties and relationships of forecast operators will allow such forecasts to be incorporated into temporal databases in a consistent way, as well as provide possible transformations for choosing the best operator for a particular application.

The remainder of this paper is organized as follows. Section 2 contains two motivating examples—one about forecasting academic grant incomes, and another about electricity connections in developing countries based on real data from the World Bank. Section 3 introduces basic notation for temporal databases. Section 4 provides an axiomatic definition of a forecast operator and then defines the classes of DF, PF and PWF forecast operators. This section also develops theorems showing relationships between DF, PF and PWF operators and the complexity of specific types of operator constructions. In Section 5, query answering mechanisms are presented that incorporate forecast operators into the standard relational algebra. Finally, related work and conclusions are given in Section 6.

2 Motivating Examples

Two motivating examples are used throughout this paper. The *grants* example specifies the total dollar amount ("Amount") of grants and number of employees ("Employees") of a Math and a CS department. Here, we are interested in predicting both of these

attributes. The *electricity* example is drawn from real World Bank[1] data about the total expenditures ("Expend") on electricity and the number of electricity connections ("Connections") in some Latin American countries. Here, we wish to forecast the number of electricity connections and the amount of total expenditures (which includes operating costs and capital investment).

	Year	Dept	Amount	Employees
t_1	2000	CS	6M	70
t_2	2001	CS	6.2M	70
t_3	2002	CS	7M	75
t_4	2003	CS	6M	75
t_5	2004	CS	7.3M	74
t_6	2005	CS	9M	80
t_7	2000	Math	1M	71
t_8	2001	Math	1.1M	74
t_9	2002	Math	1M	73
t_{10}	2003	Math	0.5M	66
t_{11}	2004	Math	1.5M	79
t_{12}	2006	Math	1.2M	77

The *grants* relation

	Year	Country	Connections	Expend
e_1	2000	Brazil	48,000,000	6.8B
e_2	2001	Brazil	50,200,000	7.5B
e_3	2002	Brazil	52,200,000	6.9B
e_4	2003	Brazil	53,800,000	6.3B
e_5	2004	Brazil	56,300,000	7.7B
e_6	2005	Brazil	57,900,000	10.7B
e_7	2000	Venezuela	4,708,215	7.7B
e_8	2001	Venezuela	4,877,084	5.2B
e_9	2002	Venezuela	4,998,433	4.3B
e_{10}	2003	Venezuela	5,106,783	3.3B
e_{11}	2004	Venezuela	5,197,020	3.1B
e_{12}	2005	Venezuela	5,392,500	3B

The *electricity* relation

3 Basic Notation

The forecasting framework discussed in this paper applies only to temporal databases. Therefore, some basic temporal database (DB) notation is introduced in this section. We assume the existence of a finite set **rel** of relation names, and a finite set **att** of attribute names, disjoint from **rel**. A *temporal relation schema* will be denoted as $\mathcal{S}(A_1, \ldots, A_{n-1}, A_T)$ where $\mathcal{S} \in$ **rel** and $A_1, \ldots, A_{n-1}, A_T \in$ **att**. Each attribute $A \in$ **att** is typed and has a domain $dom(A)$. Assume the existence of a special attribute A_T denoting time whose domain $dom(A_T)$ is the set of all integers (positive and negative). Also assume that each attribute is either a *variable* or *invariant* attribute. Invariant attributes do not change with time, while variable attributes might. In *grants*, "Dept" is an invariant attribute, while "Amount" and "Employees" are variable attributes. In *electricity* , "Country" is invariant, while "Connections" and "Expend" are variable.

A *temporal tuple* over $\mathcal{S}(A_1, \ldots, A_{n-1}, A_T)$ is a member of $dom(A_1) \times \cdots \times dom(A_{n-1}) \times dom(A_T)$. A *temporal relation instance* R over the relation schema $\mathcal{S}$ is a set of tuples over $\mathcal{S}$.

Given a tuple t over $\mathcal{S}(A_1, \ldots, A_n)$, we use $t[A_i]$ (where $i \in [1...n]$) to denote the value of attribute A_i in tuple t. We use *Attr*($\mathcal{S}$) to denote the set of all attributes in $\mathcal{S}$. Given a relation schema $\mathcal{S}$, we say that schema $\mathcal{S}_e$ is an *extension* of schema $\mathcal{S}$, denoted $\mathcal{S}_e \supseteq \mathcal{S}$ iff *Attr*($\mathcal{S}_e$) $\supseteq$ *Attr*($\mathcal{S}$).

[1] Benchmarking Data of the Electricity Distribution Sector in the Latin America and Caribbean Region 1995-2005. Available at: http://info.worldbank.org/etools/lacelectricity/home.htm

Throughout the rest of the paper, we will abuse notation and write $\mathcal{S}(A_1, \ldots, A_n)$ instead of $\mathcal{S}(A_1, \ldots, A_{n-1}, A_T)$, simply assuming that the last attribute in any schema is the time attribute.

Definition 1 (Equivalence of tuples). *Let R be a temporal relational instance R over schema $\mathcal{S}$, $\mathcal{A} \subseteq Attr(\mathcal{S})$ a set of attributes of $\mathcal{S}$, and t_1, t_2 tuples over $\mathcal{S}$. $t_1 \sim_{\mathcal{A}} t_2$ iff for each $A_i \in \mathcal{A}$, $t_1[A_i] = t_2[A_i]$. It is easy to see that $\sim_{\mathcal{A}}$ is an equivalence relation—we define a* cluster *for relation R w.r.t. the set of attributes $\mathcal{A}$ to be any equivalence class under $\sim^*_{\mathcal{A}}$, and $clusters(R, \mathcal{A})$ denotes the set of clusters of R w.r.t. $\mathcal{A}$.*

The following example shows clusters w.r.t. the *grants* and *electricity* examples.

Example 1. Consider the *grants* relation and suppose $\mathcal{A} = \{Dept\}$. Then $clusters(grants, \{Dept\})$ contains two clusters $\{t_1, \ldots, t_6\}$ and $\{t_7, \ldots, t_{12}\}$. On the other hand, if the *electricity* relation and the invariant set $\mathcal{A} = \{Country\}$ are considered, there are again two clusters ($\{e_1, \ldots, e_6\}$ and $\{e_7, \ldots, e_{12}\}$) in $clusters(electricity, \{Country\})$.

4 Forecast Operator

In this section, we formally define a generic *forecast operator* for any temporal DB and identify several families of forecast operators. Intuitively, a forecast operator must take as input some historical information and a time period for which to produce a forecast, which might include the future as well as past times where data is missing. The output of a forecast operator, however, can vary dramatically in form. For instance, forecasts can contain a single unambiguous prediction (called deterministic forecasts), or a single probabilistic forecast (called a probabilistic forecast), or a set of possible situations (called a possible worlds forecast). For each of these "types" of forecasts, the content can vary widely as well. The following definition accounts for all of these classes of forecasts, but requires that they satisfy specific desired properties.

Definition 2 (Forecast Operator). *Given a temporal relation instance R over the schema $\mathcal{S}$ and a temporal interval $\mathcal{I}$ defined over $dom(A_T)$, a* forecast operator ϕ *is a mapping from R and $\mathcal{I}$ to a set of relation instances $\{R_1, \ldots, R_n\}$ over a schema $\mathcal{S}_e \supseteq \mathcal{S}$ satisfying the following axioms:*

Axiom A1. *Every tuple in each R_i ($i \in [1..n]$) has a timestamp in $\mathcal{I}$.* This axiom says that the forecast operator only makes predictions for the time interval $\mathcal{I}$.

Axiom A2. *For each relation R_i ($i \in [1..n]$) and for each tuple $t \in R$ such that $t[A_T] \in \mathcal{I}$, there is exactly one tuple $t_i \in R_i$ such that $\forall A \in Attr(\mathcal{S})$, $t[A] = t_i[A]$.* This axiom says that tuples of R having a timestamp in $\mathcal{I}$ are preserved by the forecast operator (though they can be extended to include new attributes of the schema $\mathcal{S}_e$).

Axiom A3. *For each timestamp $ts \in \mathcal{I}$ and tuple $t \in R$, there is relation R_i with $i \in [1..n]$ containing the (forecasted) tuple t' such that $t'[A_T] = ts$ and $t' \sim_{\mathcal{A}} t$ where $\mathcal{A} \subseteq Attr(\mathcal{S})$ is a set of invariant attributes.* This axiom says that the forecasting is complete with respect to the timestamps in $\mathcal{I}$ and original tuples in R.

Note that axioms (A1) to (A3) above are not meant to be exhaustive. They represent a minimal set of conditions that any forecast operator should satisfy. Specific forecast

operators may satisfy additional properties. In addition, we reiterate that the temporal interval $\mathcal{I}$ in the above definition can represent both the future and the past, i.e., it can include times that follow and/or precede those in relation R.

Forecast operators may satisfy the following additional properties; however, they are not mandatory for definition as an operator.

Definition 3 (Coherence). *Suppose R is a temporal relational instance over a temporal relational schema $\mathcal{S}$, $\mathcal{I}$ is a temporal interval, and $\mathcal{A}$ a set of invariant attributes. A forecast operator ϕ is* coherent *w.r.t. $\mathcal{A}$ iff for each $R_i \in \phi(R, \mathcal{I}) = \{R_1, R_2, \ldots, R_n\}$, there is a bijection $\beta_i : clusters(R, \mathcal{A}) \rightarrow clusters(R_i, \mathcal{A})$ such that for each $cl \in clusters(R, \mathcal{A})$, it is the case that $\phi(cl, \mathcal{I}) = \{\beta_1(cl), \beta_2(cl), \cdots, \beta_n(cl)\}$.*

Basically, a forecast operator ϕ is coherent w.r.t. a set of attributes $\mathcal{A}$ if the result of applying ϕ on the whole relation R is equivalent to the union of the results obtained by applying ϕ on every single cluster in $clusters(R, \mathcal{A})$. For instance, consider the *electricity* example and $\mathcal{A} = \{Country\}$. In this case, a coherent forecast operator says that the number of electricity connections and the amount of expenditures in a country only depends on that country. Likewise, in the *grants* example with $\mathcal{A} = \{Dept\}$, using a coherent forecast operator implies that the amount of grants and number of employees only depend upon the department. Forecast operators are not required to be coherent because this property may not always be valid in all applications. For instance, there may be a correlation between grant amounts in the CS and Math departments (e.g., decreases in NSF funding may affect both of them proportionately). As a consequence, if the *grants* relation had an additional tuple t_{13} with information on the 2007 grant income of Math, then this may be relevant for a forecast about CS's grant income in 2007, but the coherence assumption would not allow this dependency. As such, coherence is not considered a basic forecast axiom.

Another property that forecast operators may satisfy (but are not required to) is *monotonicity*. Given a relation R, two disjoint sets $\mathcal{A}, \mathcal{B}$ of attributes[2], and two clusters $cl_1, cl_2 \in clusters(R, \mathcal{A})$, we say that $cl_1 <_{\mathcal{B}} cl_2$ iff $\forall\, t_1 \in cl_1, t_2 \in cl_2, B \in \mathcal{B}$ it is the case that $t_1[B] \leq t_2[B]$ We now use this ordering to define monotonicity.

Definition 4 (Monotonicity). *Let R be a temporal relational instance over a schema $\mathcal{S}$, $\mathcal{I}$ a temporal interval, and $\mathcal{A}, \mathcal{B} \subseteq Attr(\mathcal{S}) \setminus A_T$ two disjoint sets of attributes. A forecast operator ϕ is* monotonic *w.r.t. the pair $\langle \mathcal{A}, \mathcal{B} \rangle$ iff for each $R_i \in \phi(R, \mathcal{I})$, there is a bijection $\beta_i : clusters(R, \mathcal{A}) \rightarrow clusters(R_i, \mathcal{A})$ such that:*

- (i) *$\forall\, cl \in clusters(R, \mathcal{A})$, $cl \sim_{\mathcal{A}} \beta_i(cl)$ (i.e., $\forall\, t_1 \in cl, t_2 \in \beta_i(cl), A \in \mathcal{A}$ it is the case that $t_1[A] = t_2[A]$); and*
- (ii) *$\forall\, cl_1, cl_2 \in clusters(R, \mathcal{A})$ such that $cl_1 <_{\mathcal{B}} cl_2$, it is the case that $\beta_i(cl_1) <_{\mathcal{B}} \beta_i(cl_2)$.*

A forecast operator is monotonic if trends of attributes in $\mathcal{B}$ in the clusters w.r.t. $\mathcal{A}$ of the original relation R are preserved by the clusters w.r.t. $\mathcal{A}$ in the predicted relations $R_1, R_2, \ldots, R_n$. In the rest of this section, we study three families of forecast operators — deterministic forecasts, probabilistic forecasts, and possible world forecasts.

[2] An ordering of $Dom(B)$ for each $B \in \mathcal{B}$ is assumed.

4.1 Deterministic Forecast Operator

A deterministic forecast operator is one that returns a single relation with exactly the same schema as the input relation.

Definition 5 (Deterministic Forecast Operator). *Given a temporal relation R over the schema $\mathcal{S}$ and a temporal interval $\mathcal{I}$, a* deterministic forecast operator *(DF operator for short) δ is a forecast operator such that $\delta(R, \mathcal{I}) = \{R'\}$ with R' defined over $\mathcal{S}$.*

DF operators can be built on top of any standard time series forecast algorithm. The following example shows how simple linear regression is an instance of the class of deterministic forecast operators.

Example 2. Suppose (w.r.t. the *electricity* example) we want to forecast the amount of connections and expenditures in 2006 and 2007 using simple linear regression[3]. The function *LINREG*$(R, \mathcal{I})$ applies linear regression to each variable attribute in relation R for time interval $\mathcal{I}$. The result of *LINREG*$(electricity, [2006, 2007])$ is given below:

Year	Country	Connections	Expend
2006	Brazil	60,006,666.67	9.6B
2007	Brazil	61,989,523.81	10.157B
2006	Venezuela	5,495,630.8	1.353B
2007	Venezuela	5,623,904.6	0.473B

LINREG$(R, \mathcal{I})$ is an example of a DF operator, as it maps $electricity$ and a time interval $\mathcal{I}$ to the single relation $electricity' =$ *LINREG*$(electricity, [2006, 2007])$. In this example, *LINREG*$(R, \mathcal{I})$ also satisfies coherence w.r.t. the set $\mathcal{A} = \{Country\}$ and monotonicity w.r.t. the pair $\langle\{Country\}, \{Connections\}\rangle$.

4.2 Probabilistic Forecast Operator

Deterministic forecasts are 100% certain in their forecasts. In contrast, probabilistic forecasts also include information about the probability that a forecast is correct.

Definition 6 (Probabilistic Forecast Operator). *Given a temporal relation instance R over the schema $\mathcal{S}$ and a temporal interval $\mathcal{I}$, a* probabilistic forecast operator *(PF operator for short) μ is a forecast operator such that $\mu(R, \mathcal{I}) = \{R'\}$ with R' defined over the schema $\mathcal{S}' = Attr(\mathcal{S}) \cup \{P\}$ where $dom(P) = [0, 1]$.*

PF operators are just like DF operators except they have an additional probability attribute P. Each tuple returned by a PF operator includes the probability of that tuple being valid at the timestamp (associated with that tuple). Basically, the result of applying a PF operator can be seen as a probabilistic database [4] with tuple-level uncertainty[4]. In addition to the general axioms (A1)–(A3), we often want PF operators to satisfy a property called fact preservation.

[3] The same method shown in this example would allow us to use a variety of other traditional forecasting methods, such as logistic regression, nonlinear regression, etc.

[4] Extending the framework to the case of forecast operators dealing with attribute-level uncertainty is left as future work.

Property 1 (Fact Preservation). Let R be a temporal relational instance over $\mathcal{S}$ and $\mathcal{I}$ a temporal interval. PF operator μ *preserves facts* of R if for each tuple $t \in R$ such that $t[A_T] \in \mathcal{I}$, there is a tuple $t' \in R'$ with $R' \in \mu(R, \mathcal{I})$ such that $\forall A \in Attr(\mathcal{S})$, $t[A] = t'[A]$ and $t'[P] = 1$.

Axiom (A2) ensures that tuples having a timestamp in $\mathcal{I}$ are preserved by the forecast operator, i.e., for each tuple $t \in R$ such that $t[A_T] \in \mathcal{I}$ there is a certain tuple $t' \in R'$ such that t and t' have the same values in the attributes in $Attr(\mathcal{S})$. This property strengthens axiom (A2) for PF operators since it requires the additional condition that the probability values of the tuples in the resulting relation R' corresponding to those of R (preserved tuples) must be exactly 1.

The fact preservation property should be satisfied by a PF operator when the user trusts what is in the database; in other cases when the user does not trust the content of a database, he may choose to use a PF operator that does not guarantee fact preservation.

Example 3. Consider the *grants* relation. Suppose we want to forecast the amount of grants and employees for the CS and Math departments for 2006 and 2007, along with their probabilities. We may choose to apply a polynomial regression method *P_REG*$(R, \mathcal{A}, \mathcal{I})$, to variable attributes in each cluster in relation R w.r.t. $\mathcal{A}$ for a time interval $\mathcal{I}$. *P_REG*$(R, \mathcal{A}, \mathcal{I})$ is an operator that computes the probability that the actual value will be within one standard deviation of the forecasted value, based on a normal distribution. Assuming independence, the probability of the entire tuple is the product of the probabilities for the individual attributes.

P_REG$(R, \mathcal{A}, \mathcal{I})$ is an example of a PF operator. It first computes the forecasted values for each cluster:

Year	Dept	Amount	Employees
2006	CS	6.929471566	74
2007	CS	6.932925939	74
2006	Math	1.051905341	73
2007	Math	1.052429721	74

The probability of each forecasted value is computed as mentioned above:

CS: $P(Amount = 6.929471566 \pm \sigma | Year = 2006) = 0.68266$
$P(Amount = 6.932925939 \pm \sigma | Year = 2007) = 0.68264$
$P(Employees = 74 \pm \sigma | Year = 2006) = 0.68268$
$P(Employees = 74 \pm \sigma | Year = 2007) = 0.68268$

Math: $P(Amount = 1.051905341 \pm \sigma | Year = 2006) = 0.68268$
$P(Amount = 1.052429721 \pm \sigma | Year = 2007) = 0.68267$
$P(Employees = 73 \pm \sigma | Year = 2006) = 0.68141$
$P(Employees = 74 \pm \sigma | Year = 2007) = 0.6776$

The final relation, $grants'$ is shown below:

Year	Dept.	Amount	Employees	Prob
2006	CS	6.929471566	74	0.46604
2007	CS	6.932925939	74	0.46603
2006	Math	1.051905341	73	0.46519
2007	Math	1.052429721	74	0.46258

It is clear that every deterministic forecast can be expressed as a probabilistic forecast. Given a DF δ, a temporal relation instance R over schema $\mathcal{S}$, and a time period $\mathcal{I}$, we can define a simple probabilistic forecast operator $\mu^{simp,\delta}(R,\mathcal{I})$ to return $\{(t,1) \mid t \in R'\}$ where $\delta(R,\mathcal{I}) = \{R'\}$.

Theorem 1. *Suppose δ is a DF operator. Then, the following relationships are true:*

(i) *$\mu^{simp,\delta}$ is a probabilistic forecast operator.*
(ii) *If δ is coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. pair $\langle \mathcal{A}, \mathcal{B} \rangle$), then $\mu^{simp,\delta}$ is coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. pair $\langle \mathcal{A}, \mathcal{B} \rangle$).*
(iii) *$\mu^{simp,\delta}$ is fact-preserving.*

4.3 Possible Worlds Forecast Operator

Probabilistic forecasts still only give one value for the attributes being forecasted per time period. However, in general, there may be many possible instances of relation R at a future (or past) time point t. Possible worlds forecasts try to return not one instance as the output of a forecast, but a set of relations, each of which is a possible instance of the relation at the time being forecast.

Definition 7 (Possible Worlds Forecast Operator). *Given a temporal relation instance R over the schema $\mathcal{S}$ and a temporal interval $\mathcal{I}$, a* possible worlds forecast operator *(PWF operator for short) ω is a forecast operator such that $\omega(R,\mathcal{I}) = \{R_1, \ldots, R_n\}$ where each R_i is defined over $\mathcal{S}$ and has probability value $\mathbf{P}(R_i)$ such that (i) $\mathbf{P}(R_i) > 0$ and (ii) $\sum_{i \in [1..n]} \mathbf{P}(R_i) = 1$.*

Basically, every resulting relation instance R_i represents a possible forecasted world. Observe that axiom (A2) entails that every world includes the tuples representing facts in the temporal interval $\mathcal{I}$ that were assumed to be true in the original relation R.

Given any DF operator δ, we can define a PWF operator ω^δ. One possible method called the *discretized PWF w.r.t. δ*, denoted $\omega^{disc,\delta}$, is given below. Suppose R is a temporal relation over schema $\mathcal{S}$ and $\mathcal{I}$ is a temporal interval; $\omega^{disc,\delta}$ is defined as:

1. Let R' be the relation returned by $\delta(R,\mathcal{I})$. Consider each tuple $t \in R'$. For each variable attribute $A \in Attr(\mathcal{S})$, define $\mathbf{P}(\lfloor t[A] \rfloor) = \lceil t[A] \rceil - t[A]$ and $\mathbf{P}(\lceil t[A] \rceil) = 1 - \mathbf{P}(\lfloor t[A] \rfloor)$. The set of *tuple worlds* $tw(t)$ associated with any tuple $t \in R'$ is now defined to be:
   (a) $tw(t) = \{t' \mid$ for all variable attributes $A \in Attr(\mathcal{S})$, $t'[A] = \lfloor t[A] \rfloor$ or $t'[A] = \lceil t[A] \rceil$ and for all invariant attributes $B \in Attr(\mathcal{S})$, $t[B] = t'[B]\}$.
   (b) $tw(t) = \{t\}$ if $t[A_T] \in \mathcal{I}$.
2. The probability of a tuple $t' \in tw(t)$ is defined to be the product of the probabilities of all the variable attribute elements of t', i.e., if $X \subseteq \mathcal{S}$ is the set of all variable attributes in the schema of R, then $\mathbf{P}(t') = \Pi_{A \in X} \mathbf{P}(t'[A])$. If $tw(t)$ coincides with t, then $\mathbf{P}(t) = 1$.
3. The set of *relation worlds* $rw(\delta, R, \mathcal{I})$ is now defined to be the Cartesian product of all tuple worlds, i.e., $\Pi_{t \in \delta(R,\mathcal{I})} tw(t)$. Each member of $rw(\delta, R, \mathcal{I})$ is called a *relation world*. The probability of a given relation world $w \in rw(\delta, R, \mathcal{I})$ is given by $\mathbf{P}(w) = \Pi_{t' \in w} \mathbf{P}(t')$[5].

[5] This assumes that the events represented by different tuples in $\delta(R,\mathcal{I})$ are independent of one another.

4. Return $rw(\delta, R, \mathcal{I})$ and the probability distribution $\mathbf{P}$ on $rw(\delta, R, \mathcal{I})$.

Theorem 2. *Suppose δ is any deterministic forecast operator. Then, the following relationships are true:*

(i) $\omega^{disc,\delta}$ *is a* PWF *operator.*
(ii) *If δ is coherent w.r.t. the set of attributes $\mathcal{A}$, then $\omega^{disc,\delta}$ is coherent w.r.t. $\mathcal{A}$.*

Example 4. Let us return to the *electricity* relation and consider using the simple linear regression $LINREG(electricity, [2006, 2006])$ for just the one year 2006. The result of this operator follows immediately from Example 2 and consists of the first and third tuple in the relation $electricity'$ of Example 2. For this relation, the construction $\omega^{disc,\delta}$ creates 16 possible relation worlds. The total number of connections in Brazil in 2006 could be 60,006,666 (33%) or 60,006,667 (67%), and the corresponding number in Venezuela could be 5,495,630 (20%) or 5,495,631 (80%). The possible expenditures in Brazil are 9B (40%) or 10B (60%), and in Venezuela are 1B (64.7 %) or 2B (35.3 %). The probability of each world is the product of the probabilities of the tuples selected. As an example, for world w given below, $P(w) = (0.33 * 0.6) * (0.8 * 0.647) = 0.102$.

Year	Country	Connections	Expend
2006	Brazil	60,006,666	10B
2006	Venezuela	5,495,631	1B

It is worth noting that, as both DF and PWF operators satisfy axiom (A2), the tuples of the original relation belonging to the predicted temporal interval are preserved by DF operator δ, and then preserved by PWF operator $\omega^{disc,\delta}$ as well.

The following example shows that $\omega^{disc,\delta}$ does not preserve monotonicity.

Example 5. Assume that for countries C_1 and C_2, electricity connections are almost the same in a given year, differing only in their decimal number, as shown below:

Year	Country	Connections
2005	C_1	50,900,800.4
2005	C_2	50,900,800.8

Relation el

Year	Country	Connections
2008	C_1	50,900,802.3
2008	C_2	50,900,802.9

Relation $\delta(el, [2008, 2008])$

Suppose the result of $\delta(el, [2008, 2008])$ is the relation given above. Clearly, δ is monotonic w.r.t. the pair $\langle \{Country\}, \{Connections\} \rangle$. In contrast, $\omega^{disc,\delta}$ is not monotonic w.r.t. $\langle \{Country\}, \{Connections\} \rangle$, since there is relation world $w = \{(2008, C_1, 50, 900, 803), (2008, C_2, 50, 900, 802)\}$ in $rw(\delta, el, [2008, 2009])$ for which the number of electricity connections of C_2 is not greater than that of C_1.

The $\omega^{disc,\delta}$ construction takes exponential time to enumerate the possible relation worlds and compute the associated probability distribution; the number of tuple worlds $tw(t)$ for a tuple t is exponential in the number of variable attributes, and the total number of relation worlds is exponential in the number of tuple worlds.

Theorem 3. *Suppose R is a temporal relation instance over schema $\mathcal{S}$, $\mathcal{I}$ is a temporal interval, and $\mathcal{A} \subset Attr(\mathcal{S})$ is a set of variable attributes. For any DF operator δ, the running time of $\omega^{disc,\delta}$ is $O(2^{|\mathcal{A}| \cdot |R'|})$, where R' is the relation returned by $\delta(R, \mathcal{I})$.*

From the possible relation worlds produced by $\omega^{disc,\delta}$, a user may only be interested in examining those relations that are sufficiently probable and contain a given tuple.

Proposition 1. *Suppose R is a temporal relation instance over schema $\mathcal{S}$, $\mathcal{I}$ is a temporal interval, and δ is a polynomial-time computable DF operator. Given a tuple t over the schema $\mathcal{S}$ and probability threshold k, deciding whether there is a relation world $w \in rw(\delta, R, \mathcal{I})$ such that $t \in w$ and $\mathbf{P}(w) \geq k$ (or $\mathbf{P}(w) \leq k$) is in $PTIME$.*

Proof (Sketch). Let R' be the relation returned by $\delta(R, \mathcal{I})$. First check if there is tuple $t' \in R'$ such that by rounding its value, for each variable attribute A, we obtain t. If no, the answer to our decision problem is "no." Otherwise, keep this tuple t' and find a relation world w_{max} with max probability, i.e., $\forall t'' \in R'$, $t'' \neq t'$ create a maximal tuple world by choosing $t''[A] = argmax\ \mathbf{P}(t''[A])$ for all variable attributes A. If $\mathbf{P}(w_{max}) \geq k$, then the answer is "yes."

We can also convert a PF operator μ to a PWF operator. Two possible mechanisms are provided below where R is a temporal relation and $\mathcal{I}$ is a temporal interval:

(i) $\omega^{simp,\mu}(R, \mathcal{I})$ returns just one world as follows. Suppose $\mu(R, \mathcal{I}) = \{R'\}$. Then $\omega^{simp,\mu}(R, \mathcal{I}) = \{\pi_{Attr(\mathcal{S})}(R')\}$. In other words, it eliminates the probability column in R'. This one world has probability 1 according to the PWF $\omega^{simp,\mu}$.

(ii) $\omega^{ind,\mu}(R, \mathcal{I})$ operates as follows:
1. Compute $\mu(R, \mathcal{I}) = \{R'\}$ as above.
2. Let $\mathcal{W}$ be the power set of $\pi_{Attr(\mathcal{S})}(R')$.
3. For each tuple t in a relation $R_i \in \mathcal{W}$, let $\mathbf{P}(t)$ be the probability attribute of the tuple in R' whose non-probability attributes are identical to those of t. The probability of a particular relation R_i in $\mathcal{W}$ is set to $\mathbf{P}(R_i) = \Pi_{t \in R_i}\mathbf{P}(t) \times \Pi_{t' \in \pi_{Attr(\mathcal{S})}(R') \setminus R_i}(1 - \mathbf{P}(t'))$.
4. Let $\mathcal{W}'$ be the set of relations $R_i \in \mathcal{W}$ such that $\mathbf{P}(R_i) > 0$. Return $\mathcal{W}'$ together with the above probability distribution on this set.

The following theorem shows a strong relationship between a PF operator μ and the PWF operator $\omega^{simp,\mu}$.

Theorem 4. *Suppose μ is any PF operator. Then the following relationships are true:*

(i) *$\omega^{simp,\mu}$ is a PWF operator.*

(ii) *If μ is coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. $\langle \mathcal{A}, \mathcal{B} \rangle$), then $\omega^{simp,\mu}$ is also coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. $\langle \mathcal{A}, \mathcal{B} \rangle$).*

The above theorem holds irrespective of whether the PF operator μ is fact preserving or not. In contrast, $\omega^{ind,\mu}$ will be a PWF operator only if constructed using a fact preserving PF operator. To see this, consider a relation R containing tuple t such that its timestamp $t[A_T]$ belongs to the temporal interval $\mathcal{I}$. If PF operator $\mu(R, \mathcal{I})$ forecasts t' whose invariant attributes are identical to those of t and its probability value is $\mathbf{P}(t') < 1$, then there is a possible world returned by $\omega^{ind,\mu}$ that does not contain any tuple having invariant attributes identical to those of t. Hence, A2 would be violated.

Theorem 5. *Suppose μ is any fact preserving PF operator. Then the following relationships are true:*

(i) *$\omega^{ind,\mu}$ is a PWF operator.*
(ii) *If μ is coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. $\langle \mathcal{A}, \mathcal{B} \rangle$), then $\omega^{ind,\mu}$ is also coherent w.r.t. $\mathcal{A}$ (resp. monotonic w.r.t. $\langle \mathcal{A}, \mathcal{B} \rangle$).*

Theorem 6. *Suppose R is a temporal relation instance over schema $\mathcal{S}$ and $\mathcal{I}$ is a temporal interval. For any probabilistic operator μ, the running time complexity of $\omega^{ind,\mu}$ is $O(2^{|R'|})$, where R' is the relation returned by $\mu(R, \mathcal{I})$.*

We characterize the complexity of determining whether there is a possible world returned by $\omega^{ind,\mu}$ such that it is sufficiently probable and contains a tuple of interest t.

Proposition 2. *Suppose R is a temporal relation instance over schema $\mathcal{S}$, $\mathcal{I}$ is a temporal interval, and μ is a polynomial-time computable PF operator. Given a tuple t over the schema $\mathcal{S}$ and a probability threshold k, deciding whether there is a world w returned by $\omega^{ind,\mu}$ such that $t \in w$ and $P(w) \geq k$ (or $P(w) \leq k$) is in $PTIME$.*

Proof (Sketch). First check if $t \in \pi_S(R')$, where R' is the relation returned by $\mu(R, \mathcal{I})$. If $t \notin \pi_S(R')$, then it cannot belong to $\omega^{ind,\mu}$, thus the answer is 'no'. If $t \in \pi_S(R')$, then there is at least one possible world w that contains t. The possible world w_{max} (resp. w_{min}) that contains t is constructed using a strategy similar to that in the proof of Proposition 1. Finally, verify whether $P(w_{max}) \geq k$ (or $P(w_{min}) \leq k$).

5 Query Answering with Forecasting Operators

In this section, we study the relationship between forecast operators and standard relational algebra (RA) operators. We suggest adding new operators to the relational algebra to combine classical operators with the forecast operators presented here. Each RA operator can be augmented by forecast operators by either applying the forecast operators first and then applying the RA operator, or the other way around. Before formalizing this concept, we introduce two semantics for the evaluation of RA operators (these semantics are inspired by the notions of possible and certain answers introduced in [8]).

Definition 8 (Possibility and cautious semantics). *Given two sets of temporal relation instances S_1, S_2 whose elements are defined over the schemas $\mathcal{S}_1, \mathcal{S}_2$ respectively, and a binary relational algebra operator $op(\cdot, \cdot)$,*

(i) *the* possibility semantics *for op is the set $op^{poss}(S_1, S_2) = \bigcup_{\substack{R_1 \in S_1 \\ R_2 \in S_2}} op(R_1, R_2)$*
(ii) *the* cautious semantics *for op is the set $op^{caut}(S_1, S_2) = \bigcap_{\substack{R_1 \in S_1, \\ R_2 \in S_2}} op(R_1, R_2)$*

This definition can be straightforwardly extended to the case of unary RA operators.

Definition 9 (Forecast-first and forecast-last plans). *Given two temporal relation instances R_1, R_2 over the schemes $\mathcal{S}_1, \mathcal{S}_2$, respectively, a temporal interval $\mathcal{I}$, a forecast operator ϕ, a relational algebra operator op, and semantics $sem \in \{poss, caut\}$,*

(i) *a* forecast-first *plan is defined as*

$$\Phi_{forecast\text{–}first}(R_1, R_2, \mathcal{I}, \phi, op) = op^{sem}(\phi(R_1, \mathcal{I}), \phi(R_2, \mathcal{I}))$$

(ii) *a* forecast-last *plan is defined as*

$$\Phi_{forecast\text{–}last}(R_1, R_2, \mathcal{I}, \phi, op) = \phi(op^{sem}(\{R_1\}, \{R_2\}), \mathcal{I}) = \phi(op(R_1, R_2), \mathcal{I})$$

The latter equality in the forecast-last plan follows from the fact that $op^{sem}(\cdot, \cdot)$, with $sem \in \{poss, caut\}$, is equivalent to $op(\cdot, \cdot)$ if applied to singletons. A forecast-first plan returns a set of tuples, whereas a forecast-last plan returns a set of relations.

For some classes of forecast operators, these query policies satisfy some additional properties. The following proposition follows directly from the definition of possibility and cautious semantics for a given RA operator.

Proposition 3. *Let R_1, R_2 be temporal relation instances, $\mathcal{I}$ a temporal interval, and op an RA operator. For* DF *and* PF *operators ϕ, the forecast-first plans under possibility and cautious semantics are equivalent, that is, $op^{poss}(\phi(R_1, \mathcal{I}), \phi(R_2, \mathcal{I})) = op^{caut}(\phi(R_1, \mathcal{I}), \phi(R_2, \mathcal{I}))$.*

Depending on the particular query application, the basic forecast-first plan as given in Definition 9 can be further extended to allow for more flexibility in the forecast intervals and operators. Given temporal relation instances R_1, R_2 and RA operator *op*, then we can define the following variations of the forecast-first plan:

(i) **Multiple interval plan.** Consider two temporal intervals $\mathcal{I}_1, \mathcal{I}_2$, then a *multiple interval (MI) forecast-first* plan is defined as $\Phi_{MI}(R_1, R_2, \mathcal{I}_1, \mathcal{I}_2, \phi, op) = op(\phi(R_1, \mathcal{I}_1), \phi(R_2, \mathcal{I}_2))$. Here, two distinct forecasts are made using the intervals $\mathcal{I}_1$ and $\mathcal{I}_2$ before the RA operator *op* is applied.

(ii) **Multiple operator plan.** Given a temporal interval $\mathcal{I}$ and two forecast operators ϕ_1, ϕ_2, a *multiple operator (MO) forecast-first* plan is defined as $\Phi_{MO}(R_1, R_2, \mathcal{I}_1, \mathcal{I}_2, \phi_1, \phi_2, op) = op(\phi_1(R_1, \mathcal{I}), \phi_2(R_2, \mathcal{I}))$. In this plan, two different forecast operators are applied to the same interval, and the results are used by the RA operator.

(iii) **Hybrid plan.** Given two temporal intervals $\mathcal{I}_1, \mathcal{I}_2$ and two forecast operators ϕ_1, ϕ_2. A *hybrid forecast-first* plan is defined as $\Phi_{Hybrid}(R_1, R_2, \mathcal{I}_1, \mathcal{I}_2, \phi_1, \phi_2, op) = op(\phi_1(R_1, \mathcal{I}_1), \phi_2(R_2, \mathcal{I}_2))$. This plan combines the multiple interval and multiple operator forecast-first plans.

The remainder of this section will examine the relationships between forecast operators and some RA operators, providing results on the resulting extended relational operators that could, in principle, be used for query optimization. The result below states that, for specific kinds of selection conditions, using a forecast-first plan with possibility semantics will yield a superset of the result given by a forecast-last plan, while the cautious semantics will produce a subset.

Proposition 4. *Let R be temporal relation instance over the schema $\mathcal{S}$, $\mathcal{I}$ a temporal interval, ϕ a forecast operator coherent w.r.t. $\mathcal{A} \subseteq Attr(\mathcal{S})$, and C a selection condition filtering out whole clusters only (i.e., $\sigma_C(R) = \bigcup_{cl \in CL} cl$, where $CL \subseteq clusters(R, \mathcal{A})$). Then,*

(i) $\sigma_C^{poss}(\phi(R, \mathcal{I})) \supseteq R_i$ *where* $R_i \in \phi(\sigma_C(R), \mathcal{I})$

(ii) $\sigma_C^{caut}(\phi(R,\mathcal{I})) \subseteq R_i$ *where* $R_i \in \phi(\sigma_C(R),\mathcal{I})$

For DF and PF forecast operators, the possibility and cautious semantics coincide for forecast-first plans (Proposition 3). It then follows that, under the conditions specified above, $\sigma_C(\phi(R,\mathcal{I}))$ returns the same relation as $\phi(\sigma_C(R),\mathcal{I})$.

The interaction between forecast plans and projection RA operator is as follows.

Proposition 5. *Let R be temporal relation instance over the schema $\mathcal{S}$, $\mathcal{I}$ a temporal interval, ϕ a forecast operator, and $\mathcal{A} \subseteq Attr(\mathcal{S})$ invariant attributes of $\mathcal{S}$. Then,*

(i) $\pi_{\mathcal{A}}^{poss}(\phi(R,\mathcal{I})) \supseteq R_i$ *where* $R_i \in \phi(\pi_{\mathcal{A}}(R),\mathcal{I})$
(ii) $\pi_{\mathcal{A}}^{caut}(\phi(R,\mathcal{I})) \subseteq R_i$ *where* $R_i \in \phi(\pi_{\mathcal{A}}(R),\mathcal{I})$

Analogously to the selection RA operator, by Proposition 3 it follows that $\pi_{\mathcal{A}}(\phi(R,\mathcal{I}))$ coincides with the result of $\phi(\pi_{\mathcal{A}}(R),\mathcal{I})$ for DF or PF forecast operators. Also for the union RA operator, the relationship between forecast-first and forecast-last plans depends on the choice of possibility or cautious semantics.

Proposition 6. *Let R_1, R_2 be temporal relation instances over the schema $\mathcal{S}$, $\mathcal{I}$ a temporal interval, and ϕ a forecast operator coherent w.r.t. $\mathcal{A} \subseteq Attr(\mathcal{S})$. If $\pi_{\mathcal{A}}(R_1) \cap \pi_{\mathcal{A}}(R_2) = \emptyset$, then*

(i) $\phi(R_1,\mathcal{I}) \cup^{poss} \phi(R_2,\mathcal{I}) \supseteq R_i$ *where* $R_i \in \phi(R_1 \cup R_2,\mathcal{I})$
(ii) $\phi(R_1,\mathcal{I}) \cup^{caut} \phi(R_2,\mathcal{I}) \subseteq R_i$ *where* $R_i \in \phi(R_1 \cup R_2,\mathcal{I})$

As above, $\phi(R_1,\mathcal{I}) \cup \phi(R_2,\mathcal{I})$ is equal to $\phi(R_1 \cup R_2,\mathcal{I})$ for DF and PF operators.

6 Related Work and Conclusions

Though there are numerous works on forecasting in general [3], as well as specialized forecast models for specific domains, such as finance [15], epidemiology [9], or politics [2,10,11,14,12], all these methods vary dramatically from one another. With a large array of possible statistical models, one previous attempt to better understand the relationship between these forecasting procedures is given by [7], which integrates several forecasting methods into a common mathematical framework.

There has also been some recent work on the issue of forecasting queries in databases [1,5,6]. [5] describes the *Fa* data management system that provides support for declarative predictive queries over time series data, incorporating algorithms to effectively choose the best model type and attributes for the best query performance. Another Predictive DBMS is presented in [1] which also proposes a declarative forecasting query language, including the flexibility for both automated and user-defined predictive models. [6] investigates the I/O efficiency of forecasting queries, using a skip-list to index the time series and provide access to multiple regression models at varying levels of granularity. The model of forecasting presented in this paper differs from these prior efforts by focusing on general characteristics of forecasting rather than specific queries for a limited set of potential time series analysis methods. In fact, the framework given here can serve as a generalized, unifying theory for forecasting in databases that encompasses the semantics of these other approaches.

In this paper, we first provide axioms that any forecast operator should satisfy, together with additional desirable (but not required) properties. Our methods allow us to take classical forecasting operators and categorize forecast operators into three increasingly expressive categories and then embed them as operators in a temporal database: (i) deterministic forecast operators, (ii) probabilistic forecast operators, and (iii) possible worlds forecast operators. These classes of operators all satisfy our forecasting axioms, and in some cases, additional desirable properties. We have explored several policies for combining forecast and standard relational algebra operators to answer forecast queries and started a theoretical analysis on the interaction between these operators. Future work will focus on further investigating forecast policy w.r.t. relational algebra operators and exploiting these results as a basis for optimization of forecast queries. Though forecasting is often complex, we are able to prove that many of the techniques reported in this paper are tractable.

References

1. Akdere, M., Cetintemel, U., Riondato, M., Upfal, E., Zdonik, S.: The case for predictive database systems: Opportunities and challenges. In: Proceedings of the 5th Biennial Conference On Innovative Data Systems Research (2011)
2. Bond, J., Petroff, V., O'Brien, S., Bond, D.: Forecasting turmoil in indonesia: An application of hidden markov models. In: International Studies Association Convention, Montreal, pp. 17–21 (March 2004)
3. Bowerman, B., O'Connell, R., Koehler, A.: Forecasting, Time Series and Regression, 4th edn. Southwestern College Publishers (2004)
4. Dalvi, N.N., Suciu, D.: Management of probabilistic data: foundations and challenges. In: PODS, pp. 1–12 (2007)
5. Duan, S., Babu, S.: Processing forecasting queries. In: Proceedings of the 33rd International Conference on Very Large Databases (2007)
6. Ge, T., Zdonik, S.: A skip-list aproach for efficiently processing forecasting queries. In: Proceedings of the 34th International Conference on Very Large Databases (2008)
7. Harvey, A.C.: A unified view of statistical forecasting procedures. International Journal of Forecasting 3(3), 245–275 (1984)
8. Imielinski, T., Lipski, W.: Incomplete information in relational databases. J. ACM 31(4), 761–791 (1984)
9. Jewell, N.P.: Statistics of Epidemiology. Chapman & Hall/CRC (2003)
10. Martinez, M.V., Simari, G.I., Sliva, A., Subrahmanian, V.S.: Convex: Context vectors as a similarity-based paradigm for forecasting group behaviors. IEEE Intelligent Systems (2008)
11. Schrodt, P.: Forecasting conflict in the balkans using hidden markov models. In: Proc. American Political Science Association meetings (August 31 - September 3, 2000)
12. Sliva, A., Subrahmanian, V., Martinez, V., Simari, G.: Mathematical Methods in Counterterrorism. Springer, Heidelberg (2009)
13. Stallard, E.: Product liability forecasting for asbestos-related personal injury claims: A multidisciplinary approach. In: National Institute on Aging Conference: Demography and Epidemiology: Frontiers in Population Health and Aging, Washington, D.C (2001)
14. Subrahmanian, D., Stoll, R.: Events, patterns, and analysis. In: Programming for Peace: Computer-Aided Methods for International Conflict Resolution and Prevention. Springer, Heidelberg (2006)
15. Taylor, S.J.: Modelling Financial Time Series, 2nd edn. World Scientific Publishing Company, Singapore (2007)