Dresden Database
Systems Group

Lastgesteuerte Modellwartung für Prognoseanfragen

Lars Kegel

Defence of Diploma thesis
Supervisor: Prof. Dr.-Ing. Wolfgang Lehner
April 2nd, 2015

Motivation: Time series forecasting

Dresden Database Systems Group

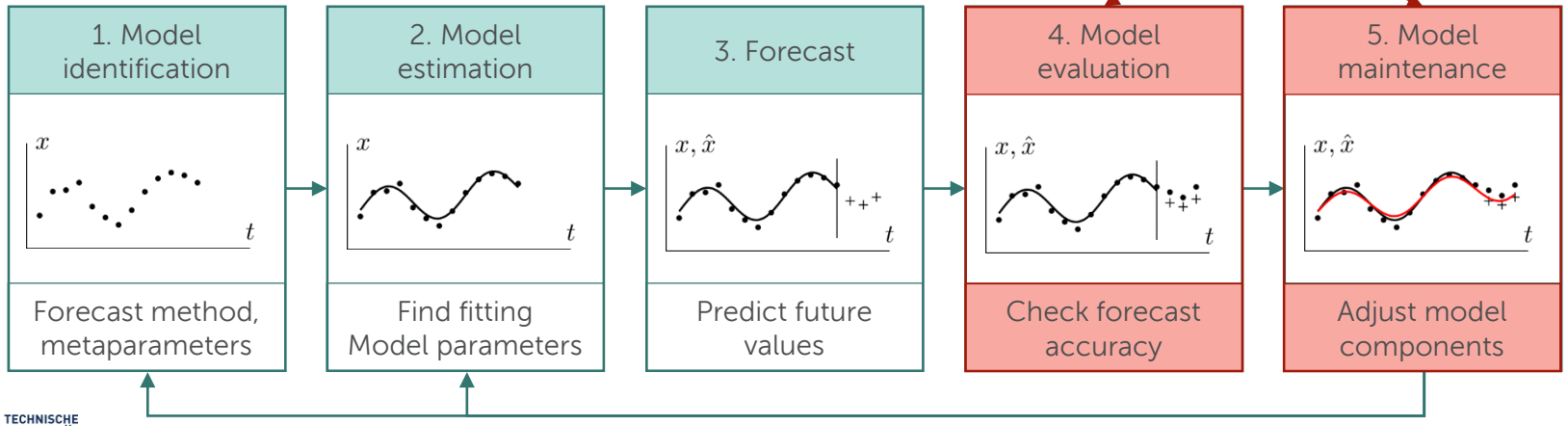
Time series forecasting

- Analysis of time series and description of the likely outcome in the immediate future
- Prerequisite for decision-making processes :
  - Production planning
  - Business intelligence
  - Energy load balancing

Components of Forecasting process

x ... Measuring value
$\hat{x}$ ... Forecast value
t ... Timestamp

The thesis is subject to these components



Motivation: Forecasting in Databases

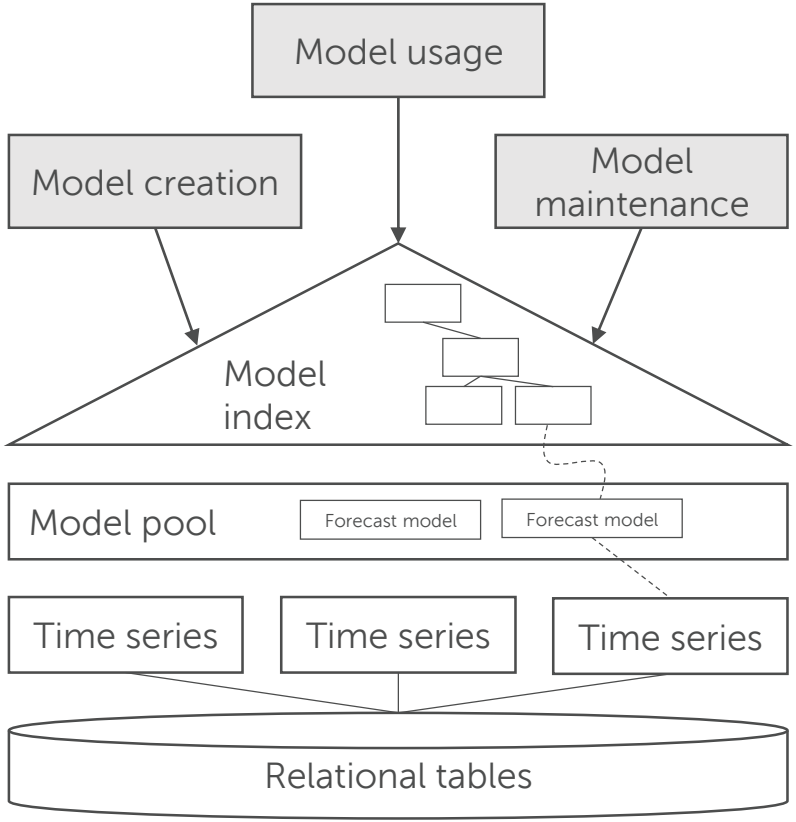
Dresden Database Systems Group

Flash-Forward Query Framework

- Integration of Time series forecasting in Database Management System (DBMS)
- Advantages
  - Time series are processed within a DBMS
  - Shared use of forecast models
  - Embed forecast in complex query

Life cycle of forecast models

- Model creation
  - Storage of forecast model in model pool
  - Indexing model for query and tuple matching
- Model usage: query matches model and returns forecast
- Model maintenance
  - Update forecast model if new tuples arrive
  - Adjust forecast model if forecast is inaccurate
  - Forecast accuracy is measured by
    - Elapsed timestamps since last model estimation
    - Forecast error since last model estimation



[Ulrike Fischer, 2014]

Motivation: Hierarchical forecasting

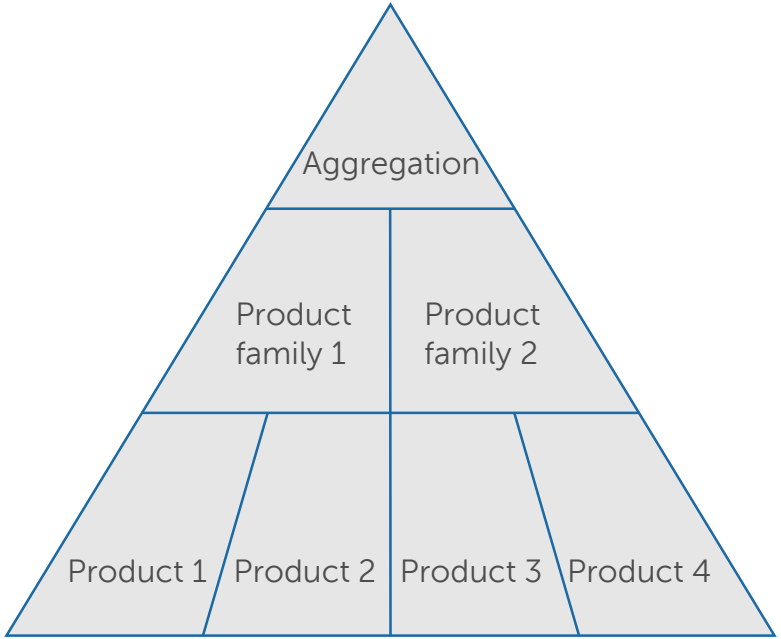
Dresden Database Systems Group

Hierarchical forecasting

- Time series may be classified by one or more hierarchies
- Figure shows functional dependency of product time series and product families
- Measuring values are aggregated to superior level
- *Forecast values can be derived, too*

Advantages

- Reduce number of forecast models when data consists of hierarchical time series
- Research shows that derivation can increase forecast accuracy



Agenda

Dresden Database Systems Group

Problem	Concept presented in this work
▪ Creation of many forecast models for a given data set	▪ Reuse of forecast models by derivation models
▪ User has different accuracy expectations on forecast models ▪ System has general accuracy assumptions only	▪ Accuracy classes for forecast queries ▪ Priorization of maintenance tasks for queried forecast models
▪ High system load causes maintenance tasks to slow down ▪ The delay of forecast queries increases	▪ Limit maintenance component to a maximum of load ▪ Provide maintenance of queried models during high system load

Derivation schemes

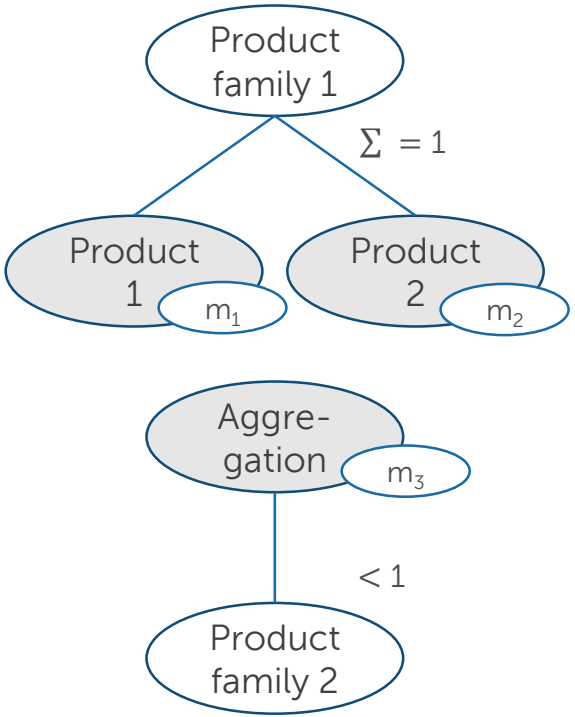
Dresden Database Systems Group

Aggregation model ("Bottom-Up")

- An aggregation model aggregates the forecast of inferior forecast models
- Example: An aggregation model for Product family 1 aggregates forecast of Products 1 and 2

Disaggregation model ("Top-Down")

- A disaggregation model uses a proration of the forecast of the superior forecast model
- Example: The forecast of Product family 2 is a proration of the aggregation, i. e. the forecast of the whole dataset
- Proration is represented by a derivation weight



Automatic retrieval of forecast models

Dresden Database Systems Group

Extension of model index

- Model index represents conditions of forecast query
- Support functional dependency in index structure
  - Functional dependency is induced by foreign key or manually (red arrows):

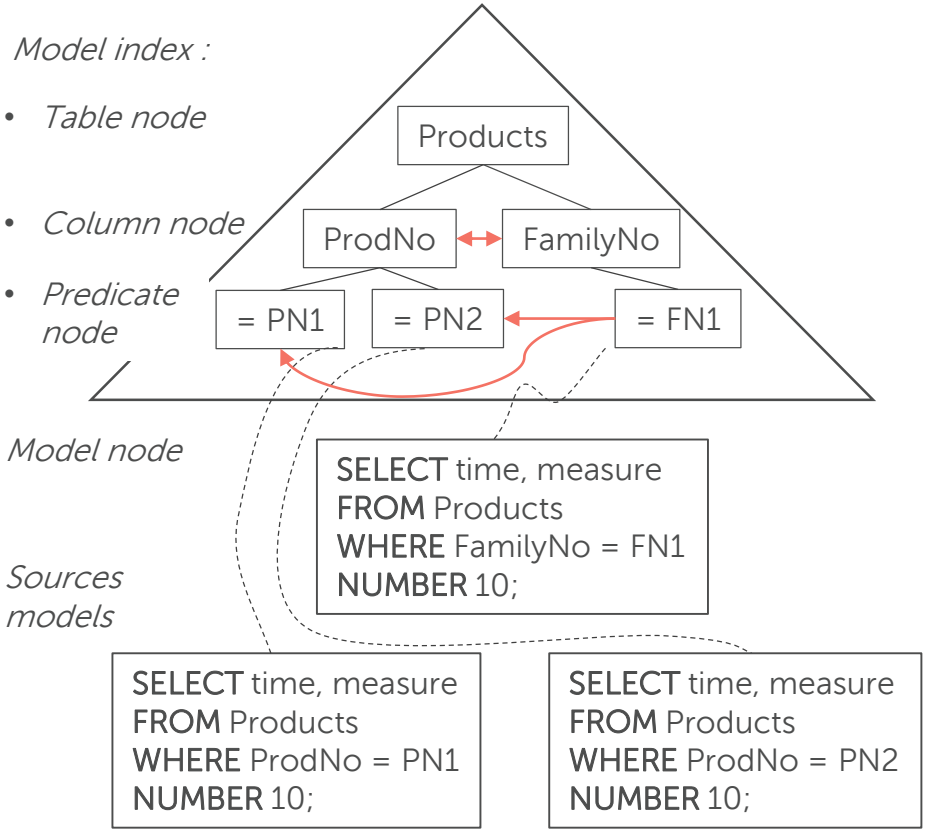
```
CREATE HIERARCHY 'Family' (
        ProdNo CHILD OF FamilyNo
) ON (SELECT * FROM Products);
```

Application to Aggregation (cf. figure)

- Find inferior dimension level and predicates
- The set of found models form the source models of an aggregation model
- Recursive: inferior model can be aggregation model

Application to Disaggregation

- Find superior dimension level and predicate
- The found model is the source model of a disaggregation model



New Maintenance tasks

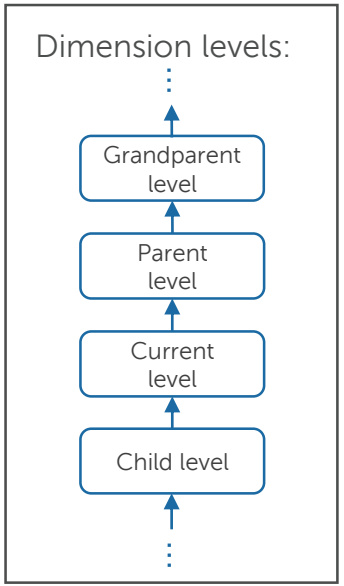
Dresden Database Systems Group

Maintenance of derivation weights

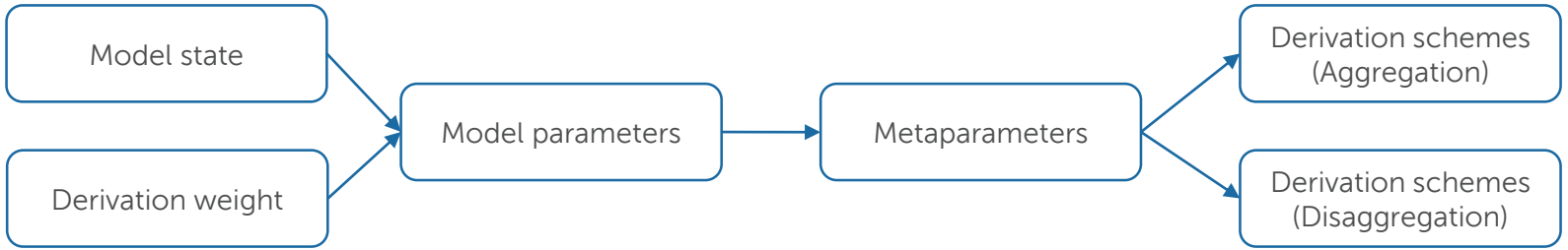
- Update of derivation weight after each new timestamp
- Task is done during tuples insert because it is fast

Maintenance of derivation schemes

- Create and evaluate derivation model of both derivation schemes in all dimensions
  - For evaluation, consider only the parent level (next higher level) and child level (next lower level)
- The derivation model with the highest accuracy is chosen for forecasting the time series
- Complex task that needs the temporary creation of source models of all derivations candidates and is time-consuming



Summary of maintenance tasks



Evaluation

Dresden Database Systems Group

Motivation

- Evaluate that derivation models can increase or maintain forecast accuracy
- Evaluate that number of forecast models can be reduced

Data set "Tourism"

- Domestic overnight trips of Australians
- Quarterly observations from 1999 to 2014
- Time series in two dimensions
  - Purpose (holiday, business, visiting friends, other)
  - Location, representing the State (New South Wales, Queensland, ...) and Region (Sydney, Blue Mountains, ...)
- 228 base time series + 112 time series for higher dimension levels = 340 time series

Procedure

- Build forecast models for 340 time series, training period 1999 to 2002
- Insert time series from 2003 to 2014
- Trigger Maintenance tasks

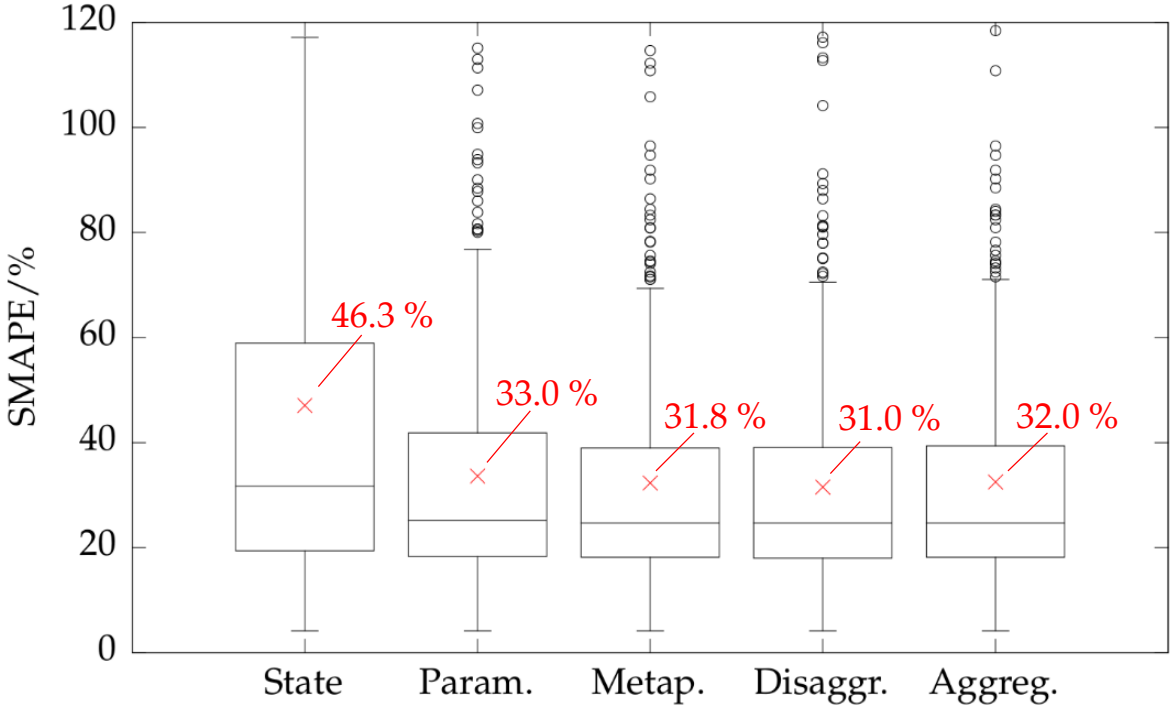
Limits	Value	Notice
Error limit	SMAPE = 24 %	-
Time limit (minimum)	4	One year
Time limit (maximum)	-	-

Evaluation

Dresden Database Systems Group

All time series from tourism

- Compare usage of maintenance tasks and forecast error on 48 measuring values



Comments

- Box plot show 5 Maintenance tasks switched on one by one
- Either Disaggregation or Aggregation is used
- Forecast error is relative between 0 % and 200 %

Conclusion

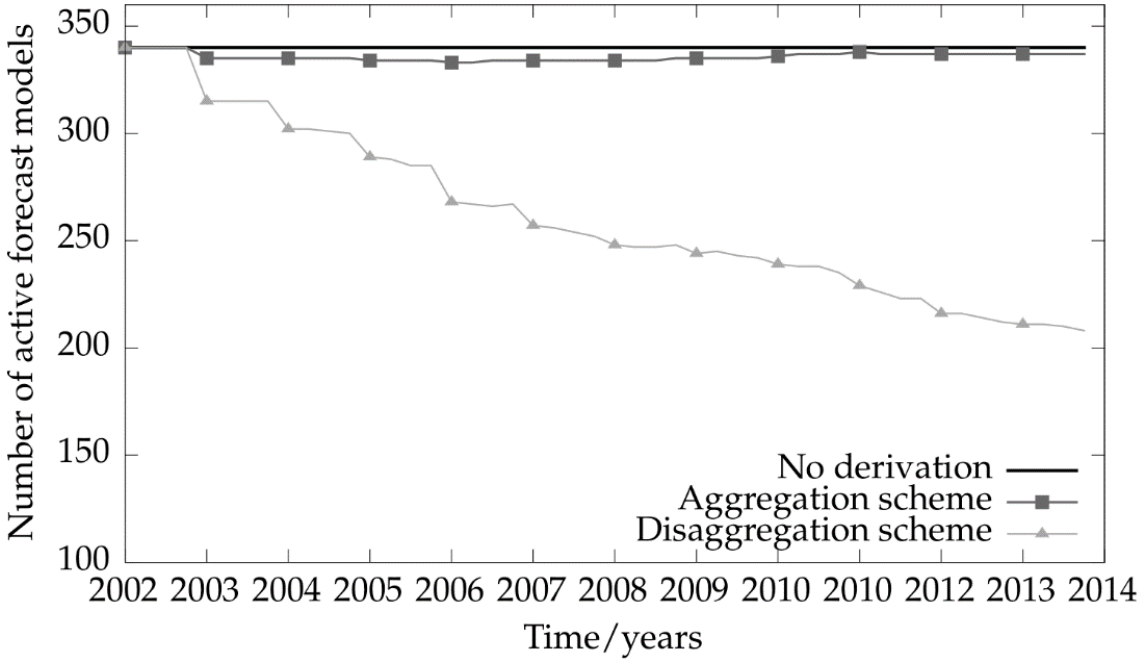
- Usage of disaggregation keeps or increases forecast accuracy
- Comparing to other maintenance tasks (modelparameter, metaparameter) the effect is rather small

Evaluation

Dresden Database Systems Group

All time series from tourism

- Measure how many forecast models are in use at each timestamp
- Show the usage of derivation schemes that reuse models



Conclusion

- Aggregation is not successful for Tourism, direct models are switched on again
- Disaggregation allows reuse of many models
- There are more timeseries which can use a disaggregation model than an aggregation model

Agenda

Dresden Database Systems Group

Problem	Concept presented in this work
▪ Creation of many forecast models for a given data set	▪ Reuse of forecast models by derivation models ✓
▪ User has different accuracy expectations on forecast models ▪ System has general accuracy assumptions only	▪ Accuracy classes for forecast queries ▪ Priorization of maintenance tasks for queried forecast models
▪ High system load causes maintenance tasks to slow down ▪ The delay of forecast queries increases	▪ Limit maintenance component to a maximum of load ▪ Provide maintenance of queried models during high system load

User-defined accuracy classes

Dresden Database Systems Group

Criteria in selecting forecasting technique [cf. Yokuma and Armstrong, 1995]

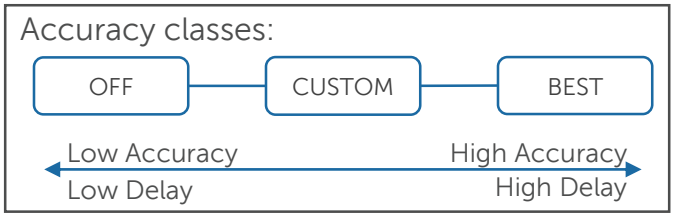
- Accuracy, cost savings from improves decisions, ease of interpretation, ease of use, maintenance cost,...
- Most important: Accuracy and Timeliness (Delay) in providing forecasts

User-defined accuracy classes

- Provide user ability to weigh most important criteria in given forecast query
- BEST: Provide best forecast accuracy that is possible but take into account a delay for maintenance tasks
- OFF: Expect fastest forecast query and ignore any maintenance task, risk of higher forecast error
- CUSTOM: User provides error limits and expected maintenance tasks

Measuring Criteria

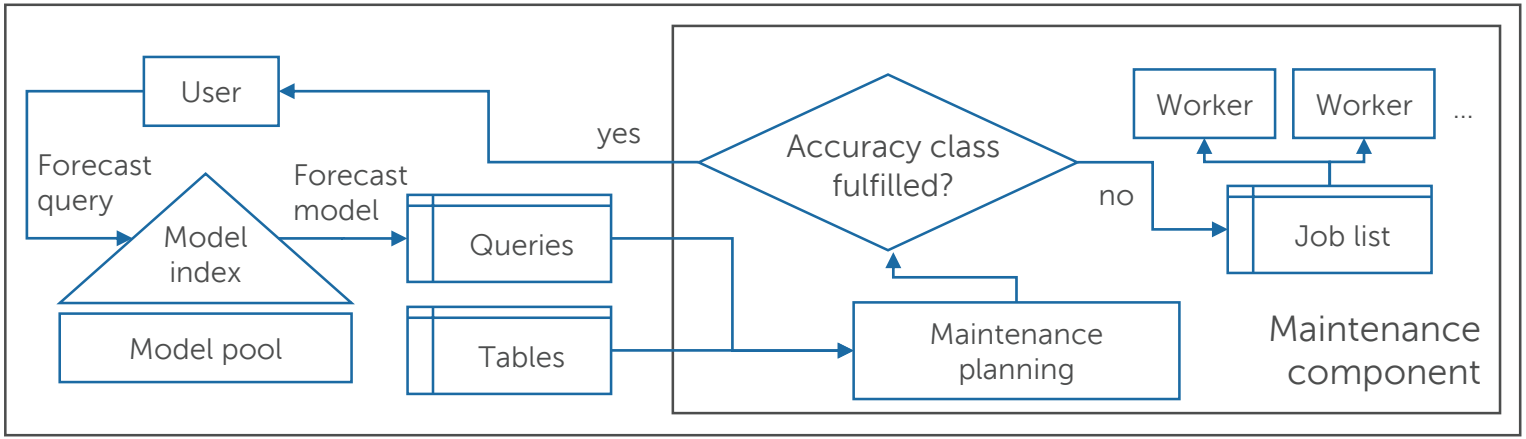
- Forecast accuracy: As before, accuracy is measured as forecast error since the last model estimation
- Timeliness in providing forecast: Query time of forecast query



Implementation

Dresden Database Systems Group

Extension of Maintenance component



Priorization of forecast queries

- During query periods, the maintenance component evaluates only queried tables
- Priorization of maintenance tasks of queried forecast models
  - Evaluation is done before evaluation of non-queried forecast models
  - Creation of a Job list with Priority that is served before Job list without priority

Evaluation

Dresden Database Systems Group

Motivation

- Compare influence of accuracy classes on forecast error and query delay
- Evaluate that CUSTOM allows gradual changes

Procedure

- Build forecast models for 340 time series of Tourism, training period 1999 to 2002
- Insert time series from 2003 to 2012, maintenance tasks supported, derivation scheme is disaggregation
- Forecast query for each model
  - Forecast accuracy on forecast 2013 to 2014, expressed as relative error measure
  - Query time is delay in maintenance component
- Increase accuracy class

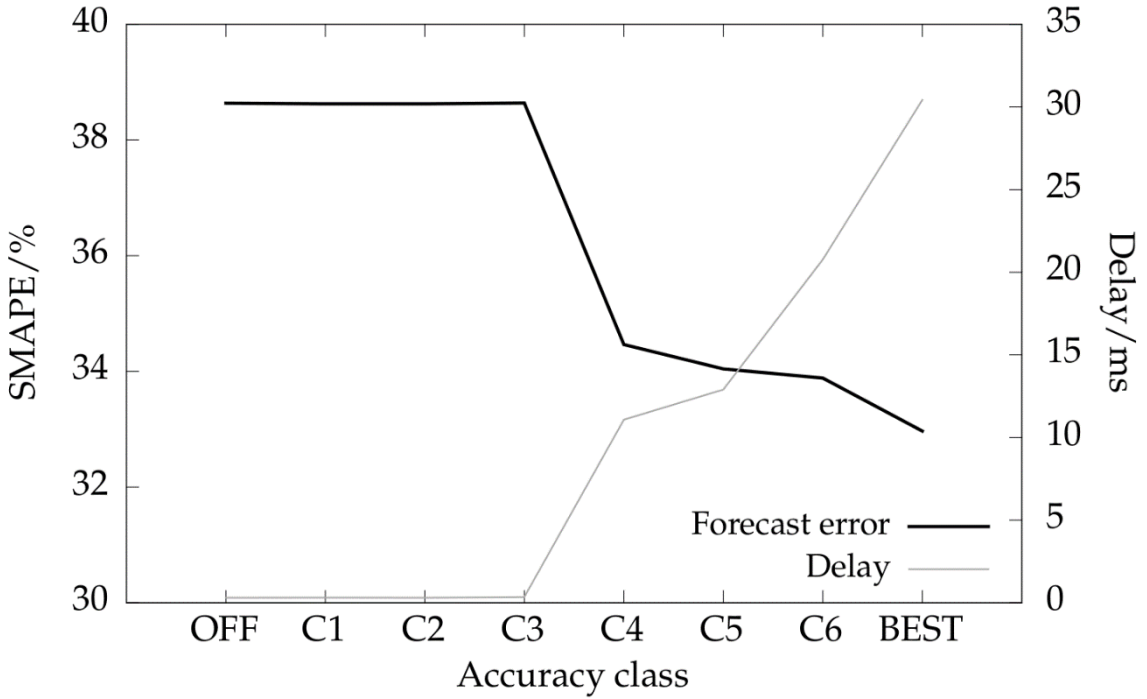
```
SELECT time, measure FROM Tourism
WHERE state = 'New South Wales'
GROUP BY time ORDER BY time NUMBER 8
MAINTENANCE CUSTOM (
        MAX_ERROR 24, MAX_TIME 16
) OPERATIONS (
        MODELPARAMETER, METAPARAMETER
);
```

Accuracy class	Highest maintenance task	Error limit / SMAPE
OFF	Model state	-
C1	Model parameter	24 %
C2	Metaparameter	24 %
C3	Derivation	24 %
C4	Model parameter	12 %
C5	Metaparameter	12 %
C6	Derivation	12 %
BEST	Derivation	-

Evaluation

Dresden Database Systems Group

Results of User-defined accuracy classes



Conclusion

- Forecast accuracy and Query time are opposed criteria, the user must weigh their importance
- Accuracy classes are more accurate than the system given error and time limits

Agenda

Dresden Database Systems Group

Problem	Concept presented in this work	
▪ Creation of many forecast models for a given data set	▪ Reuse of forecast models by derivation models	✓
▪ User has different accuracy expectations on forecast models ▪ System has general accuracy assumptions only	▪ Accuracy classes for forecast queries. ▪ Priorization of maintenance tasks for queried forecast models	✓
▪ High system load causes maintenance tasks to slow down ▪ The delay of forecast queries increases	▪ Limit maintenance component to a maximum of load ▪ Provide maintenance of queried models during high system load	

Load-controlled maintenance component

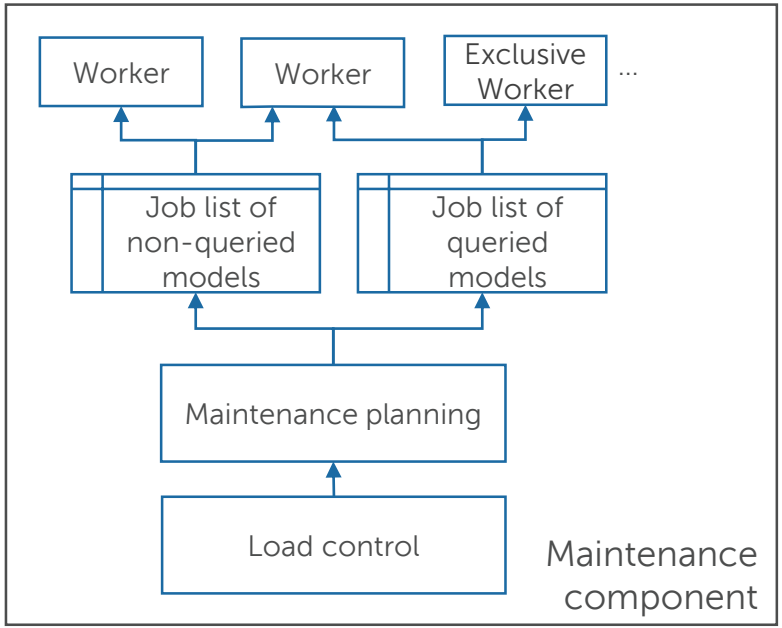
Dresden Database Systems Group

Representation of system load

- Process load represents the active clocks of each processor
- Average load represents the number of active or waiting processes in a period of time (1 minute)
- *In Unix system, Average load is a commonly used representation*

Load control in Maintenance component

- High system load is due to maintenance tasks
- During high load, limit parallel maintenance tasks by reducing number of workers for non-queried forecast models
- Exclusive workers accept jobs for queried models only and assure a fast service



Evaluation

Dresden Database Systems Group

Effects on system load

- Evaluate that controller reduces system load
- Evaluate that it sets workers exclusive

Data set "Wind"

- Energy supply data of 1326 US wind plants from 2004 to 2006 and a high resolution (10 minutes)
- Dimension Location (US states) allows derivation models
- 1326 base time series + 35 time series for higher dimension levels = 1361 time series

Procedure

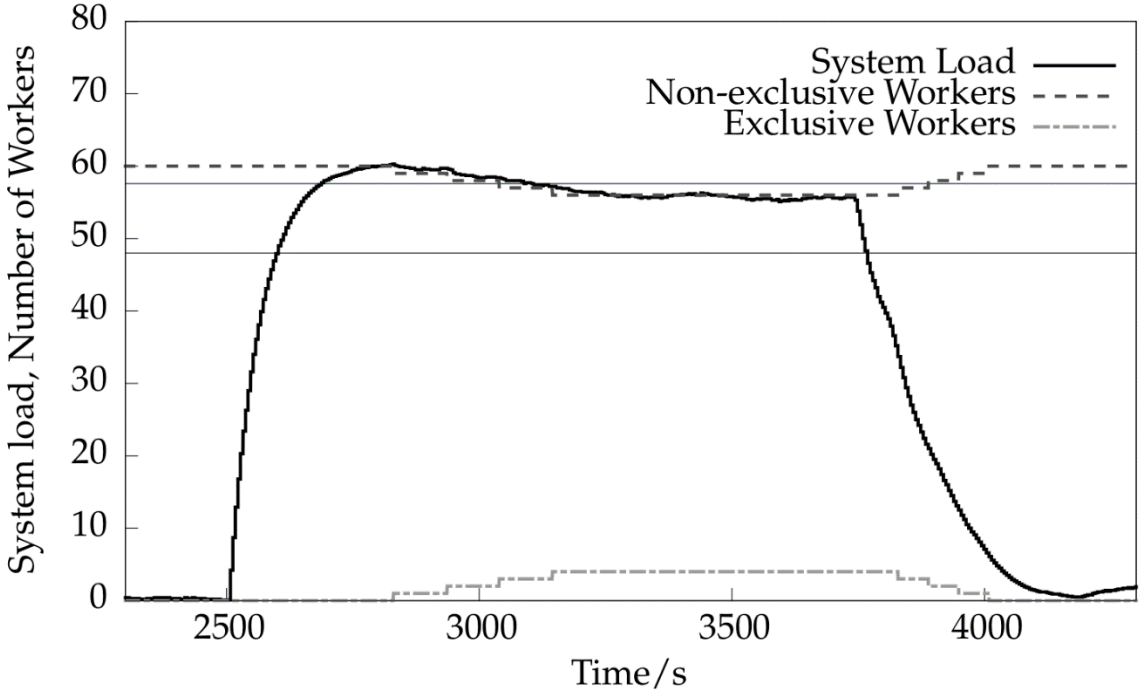
- Build forecast models for all time series, training period from 2004 to 2005
- Insert time series for January and February 2006, maintenance tasks are triggered
- Avoid hard drive access by buffering time series in main memory

Limits	Value	Notice
Error limit	SMAPE = 14 %	-
Time limit (minimum)	1008	One week
Time limit (maximum)	4320	One month

Evaluation

Dresden Database Systems Group

Effects on system load



Limits	Value
Upper limit	57.6
Lower limit	48
Delay of action	60 s

Conclusion

- During high load, the controller reduces system load under given limit
- Exclusive workers don't cause load that's why the system load is stable

Evaluation

Dresden Database Systems Group

Effects on query delay

- Create {1,10,20} users sending continuously queries to the system
- Measure query delay of forecast models that are maintained, other models pass quickly

Conclusion

- Exclusive workers allow a faster service to forecast queries during high system load

Agenda

Dresden Database Systems Group

Problem	Concept presented in this work	
▪ Creation of many forecast models for a given data set	▪ Reuse of forecast models by derivation models	✓
▪ User has different accuracy expectations on forecast models ▪ System has general accuracy assumptions only	▪ Accuracy classes for forecast queries. ▪ Priorization of maintenance tasks for queried forecast models.	✓
▪ High system load causes maintenance tasks to slow down ▪ The delay of forecast queries increases.	▪ Limit maintenance component to a maximum of load ▪ Provide maintenance of queried models during high system load	✓

Outlook

Maintenance of derivation schemes

- During evaluation of derivation models, source models have to be built temporarily
- *Avoid this time-consuming action by buffering old forecast values or by backcasting*

System load definition

- Former evaluation relies on buffering time series in main memory
- Hard drive access avoid scalability, high system load would be never reached
- *Formulation of system load considering hard drive access*

Forecast model for query rate

- In practice, the query rate is not fixed
- Analyze query rate and create a forecast model that predicts its immediate future
- *Provide exclusive workers in advance if high query rate is expected*

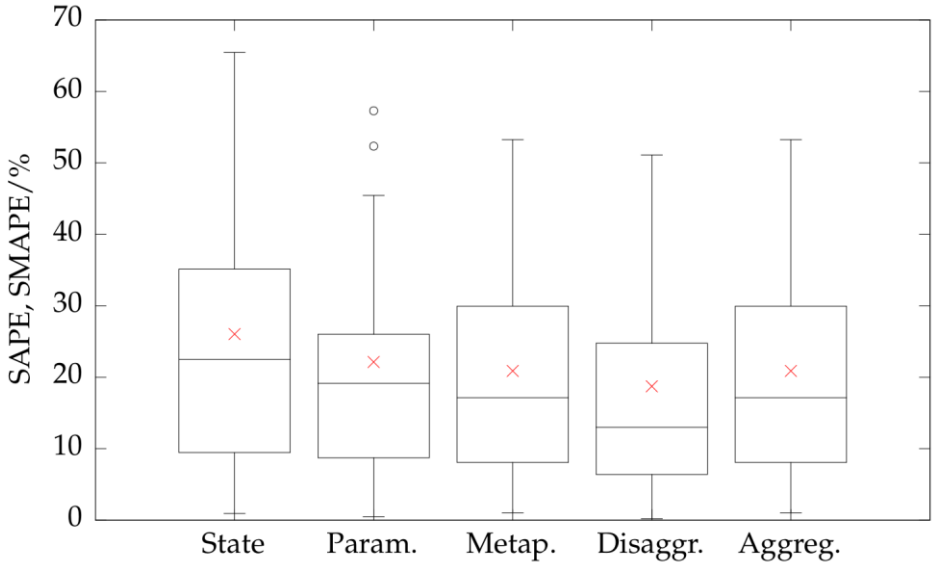
The End

Evaluation on real data

Dresden Database Systems Group

Single time series from tourism

- Compare usage of maintenance tasks and forecast error on J = 48 measuring values
- Forecast error is given as symmetric average percentage error $\text{SAPE} = 200 \cdot \frac{|x_T - \hat{x}_{T-1,1}|}{|x_T| + |\hat{x}_{T-1,1}|}$
- Mean of measuring values is given as $\text{SMAPE} = 200 \cdot \frac{1}{J} \sum_{j=1}^{J} \frac{|x_{T+j} - \hat{x}_{T+j-1,1}|}{|x_{T+j}| + |\hat{x}_{T+j-1,1}|}$



Analysis

- Decreasing mean (red cross), median, quartiles if maintenance tasks are allowed
- Usage of disaggregation model after one year
- No usage of aggregation model
- Outlier Q3 of Metap. due to wrong metaparameter estimation in 2004, correction after 1 yr

Conclusion

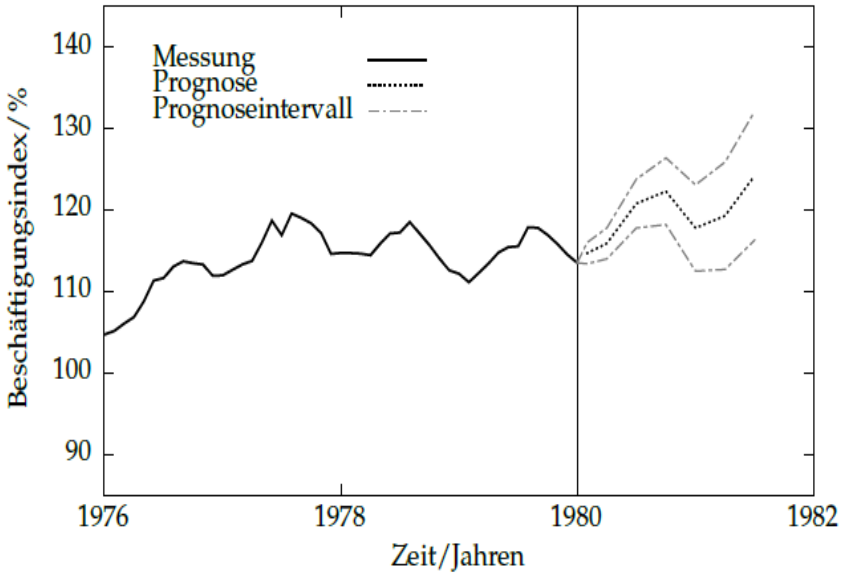
- Usage of disaggregation can increase forecast accuracy
- Derivation scheme enables to ignore own forecast model

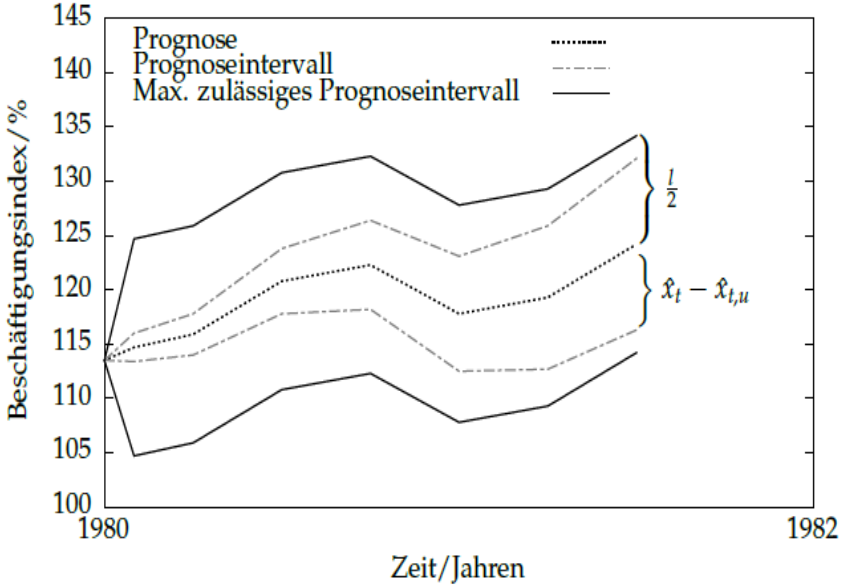
Interval forecast

Dresden Database Systems Group

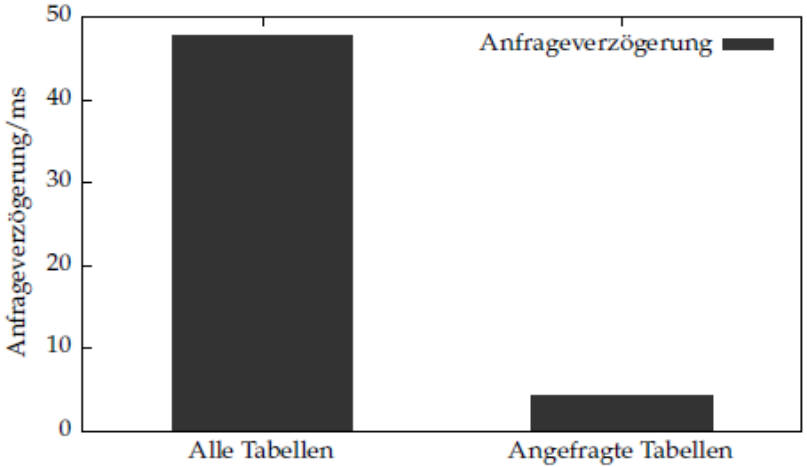
GIVE TWO PARAMETERS

- α ... Probability of forecast interval ("It's 1-α sure that the forecast value is within the interval")
- l...maximum size of forecast interval, if forecast interval is to large, it's rejected

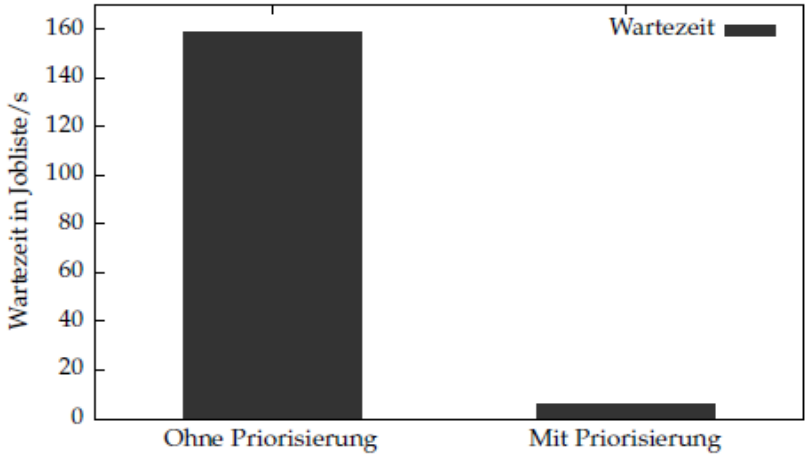




Priorization



(a) Zustandswartung angefragter Tabellen



(b) Priorisierung von Jobs

Abbildung 4.5: Minimierung der Anfrageverzögerung

Load-controlled maintenance component

Dresden Database Systems Group

Realization as three-point controller

- If current system load z exceeds upper limit z_u, then set exclusive workers
- If current system load decreases z_l, then reset workers
- Short delay time, waiting for reaction of control