Scientific workflow design: theoretical and practical issues
Terpstra, F.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)

Download date: 05 Dec 2018
Chapter 7

Data Assimilation Case Studies

7.1 Introduction

This chapter presents two case studies into the use of data assimilation for two very different applications: the prediction of bird migration and the prediction of traffic. The chapter consists of two main sections each dealing with one case study and presenting the conclusions particular to that case study. The chapter ends with more general conclusions on the lessons learned from these case studies that are relevant for implementing data assimilation as a shared software resource in a scientific workflow management system.

7.2 Bird migration model

In this case study we investigate the possibilities to predict north-south autumn bird migration over the Netherlands. The work is performed in the context of the Bird Avoidance Model, Bird Avoidance System or BAMBAS project[112]. The aim of this project is to provide accurate information to the Royal Dutch Air force on the presence of birds over the Netherlands. They can use this information to plan their training missions in such a way that the minimize the risk of bird strikes. Bird strikes to aeroplanes are a serious hazard and can cause a large amount of damage and have caused fatalities in the past. The BAMBAS project is a cooperation between multiple institutions, SOVON, University of Amsterdam and The air force. It combines scientists from different disciplines, biology, physical geography and computer science. The case study here is a first investigation into the use of data assimilation for predictive bird migration models. Results of this case study were presented at the DARE 2004 workshop [119].


7.2.1 Data

Several kinds of data are collected in the BAMBAS project which can be used for prediction of bird migration. We will briefly describe them here:

- Wind data is obtained from the KNMI the royal Dutch meteorological institute. It contains wind data over the relevant period with a resolution of 10km covering the whole of Europe and northern Africa.

- Geographical barriers include mountains such as the Alps, deserts such as the Sahara as well as seas and oceans. Birds cannot land in barriers and consequently are not able to rest or collect food in these areas.

- Vegetation/food data contains information on the amount of food available as well as the caloric value of this food.

- Radar data originates from the Dutch royal air force base in the province of Friesland. It has a maximum range of 150 km. Detailed information is available for two smaller sub samples in the radar image. Measurements are done each hour and the detailed (processed) information is based upon 10 radar rotations.

This detailed information consists of raw radar data which has been analyzed with a program that can detect birds through motion analysis. This means an estimate of speed and direction of the birds is available. The radar is not detailed enough to detect individual birds but rather groups of birds so exact bird counts are impossible\[46\]. Through the SOVON organisation bird counts of migratory birds are available\[91\]. For the collection of this data observations were made at set times of the year at 120 different locations in the Netherlands. Under the guidance of professional ornithologists, volunteers go out into the fields and for each separate area count the number of each species of bird they see within a set amount of time. This produces a record of the seasonal occurrence of migratory birds in the Netherlands.

7.2.2 Model

The model used to predict bird migration is an extended version of the model developed by Erni et. al.\[38\]. It is a model based on expert knowledge, where each bird is simulated individually. At each timestep the behavior of each bird is simulated based on the input data. Time-variable model inputs are wind speed and direction, and constant inputs are the terrain and habitat quality over Europe. The main process underlying the complex patterns of migratory flight is the bird’s energy balance. According to the terrain it is flying over and the energy it has left it can have different behaviors. The model is initialized with a population of birds leaving Scandinavia. Each bird is initialized with a different endogenous direction towards which it will try
to migrate. Currently the model is implemented in Matlab, an adaptation of an original implementation in BASIC.

7.2.3 Estimator

For the data assimilation algorithm and specifically the estimators the SOS toolbox for Matlab by Verlaan & Heemink (TU Delft)[128] was employed. Due to practical circumstances the real radar observations were unavailable for the experiments. Thus the bird observations used were generated by creating a ”truth” dataset with the Erni model. Four observations areas were defined that more or less matched the range of existing radar observations for these areas. From the ”truth” data set bird densities were derived for each timestep. White noise was added to simulate observation error. For the initial state estimate of the birds in Scandinavia no observations were available at all. Thus a guess was made consisting of a random distribution of birds over the start area in Scandinavia with bird parameters initialized according to expert knowledge.

7.2.4 Experiment

In this study real weather and terrain data were used, the radar observations were artificially generated. We first investigated the behavior of the predictions, by varying density and location of the radar observation stations (with constant, white noise observation errors). Observed Birds were simulated until the next time step, which was twelve hours in length. Every twelve hours an observation was made, after each observation the model simulated what the observed number of birds would do in the next twelve hours. At the end of the timestep an observation is made above the Netherlands and the number of birds which is predicted at this point according to the model is also noted. These two types of data are input for the estimator which adjusts the state estimate. For the following timesteps the process is repeated until all birds have migrated. The data assimilation experiment produces predicted bird densities above Netherlands.

Figure 7.1 shows our first experiment: here one very large observation point above Scandinavia is used. The number of birds observed at this point (shown in the top right graph of figure 7.1) are entered into the Erni model. This results in a predicted number of birds over the Netherlands.

Figure 7.2 shows the results of an experiment where two types of observation are combined. In this case (radar) observations from a smaller area in Scandinavia are extrapolated using bird count data to an area the size of the first experiment.

Figure 7.3 shows the effect of varying the timestep. For this experiment the observation point was placed over northern Germany, an area where a real radar observation point exists. This experiment shows that a short
Figure 7.1: One large observation point in Scandinavia. The top graph shows the number of observed birds over Scandinavia, the bottom graph shows bird numbers over the Netherlands in three ways. The solid line represents the "truth", plus signs are (radar) observations and the dashed line is the prediction produced by data assimilation.
Figure 7.2: Combining two types of observation. The top graph shows number of birds over Scandinavia in two ways: the solid line represents real number of birds, the dashed line represents extrapolated number of birds. The bottom graph shows bird numbers over the Netherlands in three ways, the solid line represents real numbers, the plus signs represent (radar) observations, the dashed line is the prediction produced by data assimilation.
Figure 7.3: The effects of varying timestep. The graphs show the number of birds over the Netherlands. The top graph has a timestep of four hours, the middle eight hours and the bottom twelve hours. The solid line represents the real number of birds, plus signs the (radar) observations and dashed line is the prediction produced by data assimilation.
time interval means that not all birds at the first observation point have the opportunity to fly to the second observation point. Furthermore the predicted values are low compared to the real and observed values. This is partly due to the fact that birds coming in over the North Sea are not observed. Another reason is that birds are initialized with an endogenous direction. This direction lies within a certain range that was kept the same, for this experiment with observation above Germany, as the first experiments with observations above Scandinavia. However birds flying due south from Scandinavia would not reach the German observation point, thus with unchanged initialization a disproportionate number of birds will predicted by the Erni model to fly south and thus not over the observation point in the Netherlands. Most importantly though the estimator does not seem to have any significant effect in adjusting the low predictions made by the model.

Figure 7.4 shows the same type of experiment as in figure 7.2. This time the observation area is over Germany which has the disadvantages described before.

Figure 7.5 shows what happens when the predictions based on observations from the North Sea are added to the predictions based on observations over northern Germany. The prediction improves slightly as it takes away one of the problems mentioned in the first experiment where only the observations above northern Germany were used, the other conditions were still applicable.

7.2.5 Conclusions bird migration

These experiments were early explorations of how data assimilation can be used with a biological model that describes behavior. While performing these experiments, reproducibility was ensured by keeping track of the methods and data used in each experiment. This was a process that was performed manually. None of the software development done in this experiment was directly reusable in the traffic prediction case study that will be described in the next section.

More interesting experiments could be to extend our analysis to the more realistic situation where the radar measurement error depends on both bird densities (especially during migration peaks) and weather conditions (especially rainfall). The trade-offs between model error - measurement error - measurement density and predictive uncertainty could be described for different weather conditions. Also, using data sources complementary to the radar observations, such as visual counts of birds at resting places, can potentially improve the initial state estimates. These observations do not suffer from state-dependent observation errors, but on the other hand they are less easily linked to the model scale. The real test for this model still has to take place, it has to be validated against real observations in order to determine its accuracy. Only after validation can truly significant
Figure 7.4: combining two types of observation. The top graph shows true number of birds (solid line) and extrapolated number of birds (dashed line). The middle graph shows the true number of birds in the radar observation area. The bottom graph shows number of birds over the Netherlands, where the solid line represents the "truth"; plus signs are observations and the dashed line is the prediction produced by data assimilation.
Figure 7.5: Combining two observations. The graphs all show number of birds above the Netherlands where the solid line represents the "truth"; plus signs are radar observations and dashed line is the prediction produced by data assimilation. The top graph is based on observations above the North Sea, the middle graph is based on observations in Germany, the bottom graph shows the two combined.
experiments be done. However as a way to gain experience in the use of data assimilation techniques the experiments proved very useful. For future experiments which will use more data sources and involve more experiment runs, keeping track of (intermediate) results and ensuring reproducibility will be increasingly difficult.

7.3 Traffic Forecasting

In this section we will look at another application of data assimilation in behavior prediction. This time the behavior being predicted is that of car drivers, more precisely the moment their behavior leads to traffic congestion. This application is presented to give insight into the current practices within data assimilation: what types of problems are encountered during setup and execution of a typical experiment. To provide the background for this research first a brief introduction will be given into the area of Intelligent Transport Systems, the area in which traffic predictions will be used. Then the specifics of prediction within ITS will be explained, including some other state of the art predictive systems. This will be followed by the description of the data assimilation solution performed in the context of this thesis. This section ends with a discussion on how this solution can benefit from an e-Science solution and what the additional requirements for such a solution are. Results of this case study were presented at the ITS world congress 2006[118].

7.3.1 Intelligent Transport Systems

Since the beginning of the industrial revolution, ever more complex transportation infrastructure has emerged, starting with mass transportation using trains and getting even more complex with the adoption of the car as a method for personal transportation. To keep this infrastructure working and accommodate an ever increasing number of travelers requires smart solutions. Intelligent Transport Systems (ITS) try to provide technologically enabled solutions to these transportation challenges. ITS concern themselves with three main areas:

- Information
- Management
- Automation

The area of information contains such subjects as providing accurate information to travelers. This can be information on congestion or delays in public transport. It can take many forms: websites, roadside displays, mobile services etc.
Management in ITS deals with keeping traffic moving. This can take the form of traffic management systems that monitor the situation on the road through various sensors and take appropriate action when a problem is detected. For instance closing a lane when an accident is detected. It also includes systems to control trafficflow such as variable speed limits or traffic lights at highway access ramps. Probably the most well known subjects in this area are electronic toll collection systems such as “Rekeningrijden”, “Kilometerheffing” in the Netherlands and the congestion charge in London.

Automation concerns itself with assisting or fully automating the driver of a vehicle. An example of assistance is adaptive cruise control where the cruise control automatically reduces a vehicles speed when the distance to the vehicle in front gets to small. Prototypes for automated highway systems, where cars drive themselves totally autonomously, also exist. But they are still far from widespread adoption.

In the area of providing information as well as in traffic management being able to accurately predict where congestion is going to occur is a great asset. The traffic situation on the roads is a constantly changing process: thus a stretch of road which is free from congestion at the time the information is provided can be congested by the time a driver arrives there. Similarly in traffic management prediction is instrumental in taking preventative measures and also in knowing whether a certain traffic management decision will have the desired effect.

### 7.3.2 Prediction for ITS

The ITS domain has its specific properties and demands where prediction is concerned. What follows is an elaboration on these properties and demands with regard to the data, the modeling techniques employed and finally the prediction itself.

**Data**

There are many different ways of obtaining traffic data. Almost all of them fall into two main categories:

- Fixed sensors
- Floating car

Important measurements are:

- Traffic density: cars per time unit at one location.
- Speed
- Traveltime: time needed to travel a route.
Fixed sensors such as induction loops in the road and cameras can observe cars at one location. This results in accurate measurements of both the speed as well as the density of traffic. The downside is that measuring every road in this way is very expensive, which is why these types of measurements are mainly done on busy highways.

Floating car data, where the speed and position of a few cars in the traffic-flow is measured through some form of transponder, offers the advantage that traffic on all roads can be measured. The downside of floating car data is that deriving an accurate density is more difficult. However when enough measurements are available traffic speed can be estimated relatively accurately. Furthermore previous research into modeling traffic has shown that there is a relation between the speed of traffic, the capacity of the road and the density of traffic [73]. Thus, knowing the capacity of a road and the speed of traffic traveling on it, an estimate of the density can also be made. For modeling larger areas or long time spans, origin destination matrices are commonly used data. These matrices contain data on where cars start their journey, on what time they start and what their destination is. This data can be estimated through the combination of traffic measurements from multiple locations, but also by conducting interviews with motorists on their driving habits.

Cars themselves are not the only measurable thing that can be of importance when predicting traffic. Other factors exist that influence the behavior of drivers[123]. These factors are divided into three distinct categories.

- First there is nature, which in this case is almost entirely accounted for by the weather. Different weather conditions such as rain, fog and snow cause different driver behavior and therefore different traffic patterns.

- Second there are the human causes, consisting of two sub-categories of global and local influence. The global events are calendar related phenomenon such as the difference between weekends and work-days but also less regular events such as holidays. These events have an effect on the society or country in which this calendar is used, unlike the first category which is bound to one geographical location. Large public events such as pop festivals, parades and most commonly roadworks, can cause localized increases in traffic.

- The third category are incidents and accidents, which have a bidirectional relationship with traffic patterns. They cause extra congestion and capacity reduction, while creating traffic patterns that influence the chance of accidents occurring. Incidents and accidents are a stochastic process where one small event sometimes has large consequences for overall traffic patterns.
Modeling

Within ITS and traffic research in general three different scales of traffic model exist.

- Microscopic
- Mesoscopic
- Macroscopic

Microscopic models model the behavior of each driver individually and are therefore very detailed. Behavior in this case can be: when a driver changes lane, how much distance is kept to the car ahead and at what time a driver brakes to preserve this distance.

Macroscopic models on the other hand regard the traffic stream as a whole and model it as if it was a gas or liquid. This offers less detail but is also less computationally intensive. In [73] macroscopic models are explained in great detail.

Mesoscopic models are hybrids between micro- and macroscopic models. They use a macroscopic approach to straight roads but revert to microscopic methods at more complicated areas such as junctions. An example of this type of modeling can be found in the DynaMit project [42]. For a slightly dated but very comprehensive overview of traffic models one can look at the results of the "smartest” project.$^1$

Modeling factors which influence traffic can be done in multiple ways. Within microscopic models different behaviors for varying weather conditions can be modeled, while parameters in macroscopic models can be adjusted to account for the different dynamic of traffic under varying weather conditions.

Prediction

Choosing the right type of model for a particular application can depend on multiple factors. In case of traffic prediction the size and complexity of the road network. These factors determine what type of models are computationally feasible.

Another important factor is the amount of time ahead a prediction is required to be. If a prediction is further into the future it is less likely that a predicted traffic situation at a certain point can be based on direct measurements. If measurements of the current situation of the traffic situation are taken as the basis for a model that propagates the movements of measured cars into the future, then at some point in the future all cars will have disappeared from the area of interest. The simple reason behind this is that

$^1$http://www.its.leeds.ac.uk/projects/smartest/
traffic is always modeled as an open system. The problem we are dealing with here is known as the courant condition[54].

This is a problem for forecasts with a larger scope in time, or for a small modeled road network. To overcome this problem the factors influencing traffic mentioned earlier in this section need to be taken into account when constructing a model. Especially the social factors determine how many cars are going to enter and leave the network and therefore have to be included in the model to compensate for the lack of relevant direct observations. It is also important to use the available direct observations to correct the estimate of traffic based on social factors; an application that lends itself to data assimilation.

7.3.3 Current solution

To get a better idea of how data assimilation can currently be used in the ITS field we look at a suitable case study concerned with longer term traffic prediction. In this case direct observations of traffic do not provide enough information for prediction and social factors have to be modeled as well. First the available data will be described in detail. The modeling effort for this case study involves two methods. First Auto Regressive Integrated Moving Average (ARIMA) will be introduced, a common technique in time series analysis which will be used as a baseline to compare to the Dynamic Harmonic Regression (DHR) method. DHR was chosen because it has been used successfully in comparable applications, both involving longer term prediction[125] and involving shorter term traffic prediction[117]. Finally the results of the experiment are analyzed.

Data

For this case study we use data concerning one 500 meter stretch of highway. This allows for (future) comparison with detection loop data. The data covers the whole of October 2004. It includes three normal working weeks and a one week school holiday. The raw data consists of speed measurements at a one minute granularity. However since we are interested in longer term predictions the data is first aggregated to one hour data.

As a first step towards building a predictive system we will look at properties of the data that will have an impact on any modeling attempt. First we looked at the autocorrelation, to see which cycles, if any, were present in the data. In figure 7.6 one can see correlations for 24 hours and all subsequent multiples of 24 which decay slowly with each multiple except for the 168 hours point (one week), where it is a lot stronger. The correlation at one week is the most interesting, as it implies that not all days of the week have similar traffic patterns. Furthermore there are strong negative correlations to be seen at 12 hours after the 24 hour peaks. This could be explained as an
imbalance between morning and evening rush hour. This is very plausible since the data concern only one direction of traffic flow where one would expect just one rush hour per day. As a next step we looked at the correlation between individual days to determine for which day this weekly correlation was strongest. In figure 7.7 we show the correlation between all Mondays in October 2004. This is achieved by taking the traffic pattern of each Monday in October and calculating the correlation coefficient for all possible pairs of Mondays. It can be seen that Mondays are strongly correlated. Fridays, although not shown in these figures, produce results similar to Mondays, also showing a strong correlation. Days in the middle of the working week, such as Thursdays are much less correlated as can be seen in figure 7.8, which was calculated in the same manner as figure 7.7. On the other hand these midweek days (Tuesdays, Wednesdays and Thursdays) show much stronger one and two day correlations. This is demonstrated by figure 7.9 where the correlation between three consecutive midweek days is shown.

Finally Saturdays and Sundays do not show much correlation at all. They don’t have a weekly correlation nor a correlation with the other consecutive day in the weekend nor with any other day of the week. The explanation for this is twofold. Firstly in weekends there is much less traffic than in other days of the week. This does not result in regular traffic jams, which form the basis of the weekday correlations. Secondly, because there is less traffic, there are far less observations making the resulting traffic patterns less accurate and more erratic, even when aggregated over 1 hour intervals.

There are several lessons to be learned from the data analysis presented above, with regard to modeling. Weekends are, with the current data set, best ignored. Our main interest is in predicting congestion with a certain degree of reliability. Including data from weekends would do more harm
Figure 7.7: Correlation coefficients for the pattern of traffic speeds on the four Mondays of October 2004
Figure 7.8: Correlation coefficients for the pattern of traffic speeds on the four Thursdays of October 2004
Figure 7.9: Correlation coefficients for the pattern of traffic speeds on the first Tuesday, Wednesday and Thursday of October 2004
7.3. TRAFFIC FORECASTING

than good to an overall prediction because of a lack of accuracy and the absence of congestion in the data. The correlation between different weekdays suggest the possibility of modeling midweek days, Mondays and Fridays as three separate entities. Lastly, when evaluating the results of forecasts we should not forget the school holiday in the last week of our dataset. Earlier in this section calender based events were given as an instance of human factors of influence on traffic. This influence is noticeable in the dataset through a reduction of congestion during the holiday period and also by the distinct traffic patterns of different weekdays, especially Mondays and Fridays. It should be noted that the correlation even for these weekdays is quite weak (below 0.85). However, the difference between Mondays, Fridays and weekdays for traffic data is well known within traffic research. Other studies have made the same categorization based on their traffic data[69] as well.

Dynamic Harmonic Regression

The modeling method chosen for this traffic prediction application was Dynamic Harmonic Regression (DHR)[132], one of the methods available in the Captain toolbox[133] that was introduced in the previous chapter. The Captain toolbox has been used for all kinds of time series analysis, including traffic prediction[117]. DHR will now be explained in more detail, followed by an explanation of ARIMA(not part of Captain) which was used as a baseline method to compare it with. DHR can be used for forecasting, backcasting, interpolation over gaps in the data, signal extraction and adaptive seasonal adjustment. In the past it has been successfully used in another time-series application to forecast telephone call patterns up to several weeks ahead [125]. The unobserved components within the DHR model are the trend, seasonal and cyclical components. These hidden components will match with some of the calendar based influences on traffic mentioned in section 7.3.2. Each of the hidden components has Time Variable Parameters (TVP) associated with them to deal with non-stationarity: for instance changes of phase and amplitude over time present within a time series. A Kalman filter is used to optimize these TVP’s thereby minimizing model error. Thus if a DHR model is used for prediction at each time step we in effect have a data assimilation system minimizing model error. To model traffic using DHR and other tools in the captain toolbox, the following steps are performed. First the data is prepared, the granularity of the model has to be chosen. Observations have to be aggregated into discrete timesteps. These timesteps have to be short enough to capture the shortest (in time) traffic jam but they should not fall outside the limits allowed by the method of observation. In this case five minute timesteps were chosen. If gaps exist in the data they can be filled in. In this case the method of observation is such that a lack of data suggests there is no congestion. Thus traffic can
be assumed to be at or near the maximum speed allowed for the stretch of road concerned. Next the data is analyzed. The Akaike Information Criterion (AIC) as implemented within the Captain toolbox, can be used to find a good compromise between model fit and model complexity. From the AIC output an appropriate number of Auto regressive parameters can be chosen for use in DHR. Further analysis can be provided by viewing the auto regressive spectrum of the data through the "arspec" (auto regressive spectrum) function. What 'arspec' does in the frequency domain, the "period" function can do in the time domain. This provides evidence of cycles in the data, and verifies whether expected cycles are present. The period of these cycles is used as an input for DHR. Other inputs for DHR such as the optimization method of the TVP's have to be selected on the basis of the knowledge gained from either the manual, expert advice or from previous experience. The modeling for DHR consists of two steps. First the initial Noise Variance ratio (NVR) of the hyper parameters is determined using the "dhropt" (dynamic harmonic regression optimization) function. These hyper parameters determine the statistical properties of the Time Variable Parameters. Then the DHR function itself is run to perform the actual modeling and any fore- or backcasting which may be called for.

**ARIMA**

The name of the model used as a baseline, ARIMA, is an abbreviation for Auto Regressive Integrated Moving Average. It models a time series by
first removing trend and periodic components from the data, then fitting an ARMA model [48], and for the final result adding, or 'integrating', the previously removed trend and periodic components. Predictions are made by applying a transfer function to the model that is fitted to the known data. For more details on ARIMA see [48]. ARIMA is a very well known but relatively basic technique. If DHR is used correctly and is suitable for the domain of traffic prediction it should produce significantly better results than ARIMA.

Results

Having described the DHR and its associated methodology we will now look at the actual results of experiments using DHR. We start with preparation of the data. Weekends contained very little useful information, because very few measurements are made during that period. They were ignored for the experiments because of the aforementioned reason and because weekends contain few to no traffic jams in any case and thus are not of interest in predicting traffic jams. The weekly correlation that is mainly present in Mondays and Fridays is not modeled well when using the complete data set. Thus for the results presented below we separated the data set into three different sets, for Mondays, Fridays and midweek days respectively. The data used for predicting Mondays only consists of Mondays, the one for Fridays only consists of Fridays while the data for predicting midweek days consists of Tuesdays, Wednesdays and Thursdays. As a first experiment traffic speeds were predicted for Tuesday the 12th of October. The results in figures 7.11 and 7.14 show that big slowdowns indicating traffic jams are absent. Both methods do well, but DHR is slightly closer to the observed data. A more interesting example is shown in figures 7.12 and 7.15. This shows a twenty-four hour prediction for Friday the 22nd of October. For this day a traffic jam is clearly visible in the data with a peak at 17:00. The prediction from the ARIMA model has little relation with the true data. DHR on the other hand shows more promise, its prediction contains a traffic jam of the same amplitude, however the phase is shifted resulting in a prediction that is three hours early. Finally in figures 7.13 and 7.16 a sixteen hour prediction is shown where the first 8 hours of Friday the 22nd were also included in data on which the prediction is based. This shows that the difference between the predicted traffic jam and the actual traffic jam has become less for DHR but with ARIMA no clear prediction of a traffic jam can be observed. It is clear from the previous section that DHR shows promise as a modeling method, outperforming ARIMA especially when it comes to predicting traffic jams. These experiments have been a first step towards a traffic forecasting system, however further improvements are possible and needed.

In this experiment Data assimilation was used to minimize prediction
Figure 7.11: Twenty-four hour prediction for Tuesday 12-10-2004 using ARIMA model with midweek data from the previous week.
Figure 7.12: Twenty-four hour prediction for Friday 22-10-2004 using ARIMA model with data from three previous Fridays
Figure 7.13: Sixteen hour prediction for the last sixteen hours Friday 22-10-2004 using ARIMA model with data from the first eight hours and three previous Fridays.
Figure 7.14: Twenty-four hour prediction for Tuesday 12-10-2004 using DHR model with midweek data from the previous week.
Figure 7.15: Twenty-four hour prediction for Friday 22-10-2004 using DHR model with data from three previous Fridays
Figure 7.16: Sixteen hour prediction for the last sixteen hours Friday 22-10-2004 using DHR model with data from the first eight hours and three previous Fridays
error, that is to try to minimize the error inherent in the modeling method. Data assimilation can also be employed to minimize the observational error. To do this, knowledge about the way data is collected has to be modeled as well. For instance the number of observations behind an input value for the predictive model should be taken into account. Especially during nighttime when few observations are made, a few unusual observations can lead to a big distortion. The predictive model itself could be improved by switching to a multi variate approach where one or more of the external influences mentioned in section 7.3.2 is included.

7.3.4 Discussion and Requirements

The current solution is implemented entirely inside Matlab, it could benefit from a move to an e-Science environment. The methodology involved in using DHR is clearly suitable for implementation as a workflow. Benefits to this approach will be greater in particular areas. Increased computational power through the use of the grid is an area which offers relatively little potential improvements. In its current form additional computational power is not needed. The model is not very computationally intensive. The two scenarios in which grid computing could become useful are first when modeling a far bigger road network, or secondly when taking a brute force approach to analyzing the best model parameters. Both of these cases require many different instantiations of the model to be run, a task that is “embarrassingly parallel” and thus very well suited to grid computing. There are however far better methods for choosing the right parameters with the Captain toolbox. The toolbox offers analysis tools and an extensive manual which aid in finding the right parameters. Further knowledge is contained in papers concerning the individual tools of the toolbox[132]. The greater benefits of using an e-Science environment in this case can be found in the availability and sharing of knowledge. Easy access to the “provenance” of the workflows, data and tools involved in the experiment will help in disseminating implicit knowledge on how to properly use DHR in a traffic application. Knowledge can also be made available explicitly, in the simplest case by making existing documentation available in an on-line form, through a help function for each tool for instance. Expert knowledge on methodology can also be explicitly shared in a more interactive form. An expert in DHR can make template workflows that describe common ways of combining tools from Captain. An expert in traffic trying to employ the Captain toolbox in an e-Science environment can use both implicit and explicit knowledge to more efficiently create his experiment. Finally the specific dataset used in this experiment places demands on use in an e-Science environment. The data is stored on a remote database and as it originates from a commercial party the data has to be kept secure. Access to this data needs to be seamless to someone using it in an e-Science environment, but at the same time its confidentiality
has to be ensured. Only users with permission should be able to access it. Protection has to be robust enough for the commercial party to trust that access to its data will only happen under the terms that were agreed upon.

### 7.3.5 Conclusions for traffic prediction

We have looked at how we can enable accurate travel advice based on the prediction of the future state of traffic. We described the effect of the calendar on traffic and showed how modeling these effects explicitly can result in sensible predictions. We presented a data assimilation approach to creating a forecasting system. Furthermore we have experimentally shown the use of Dynamic Harmonic Regression for traffic forecasting and described how Dynamic Harmonic Regression fits within the data assimilation approach. Based on the results of the experiments we have discussed how the traffic forecasting system can be further developed within the data assimilation approach. The last ten years of research in Traffic Information Systems has shown that it is difficult to create useful services, unless the services that are offered have a high degree of accuracy and consistency. We are aware that experimental validation is key to this type of research.

### 7.4 Conclusion

The two case studies presented in this chapter generate very similar requirements if they are to be implemented in an e-Science environment. Both would benefit from features that ensure reproducibility, such as provenance recording, both for the data used and produced in experiments as well as for computational elements which comprise the experiment. The second benefit are features which help in the composition of experiments. In this regard there is a difference between the two case studies in that the SOS toolkit used for bird migration is relatively minimal in making explicit knowledge about the way it should be used where as the Captain toolbox is as well documented as one can expect for a stand alone toolbox. Yet even in the Case of the Captain toolbox much more knowledge about its use can and should be made explicit for use as a shared software resource in an e-Science environment. The documentation for Captain is focussed on the use of individual tools rather than the methodology needed to combine them all. The next chapter will show how such a methodology can be supported through a workflow in an e-Science environment.