

InfraWatch: Data management of large systems for monitoring infrastructural performance

Arno Knobbe¹, Hendrik Blockeel^{1,2}, Arne Koopman¹, Toon Calders³, Bas Obladen⁴, Carlos Bosma⁴, Hessel Galenkamp⁴, Eddy Koenders⁵, and Joost Kok¹

¹ LIACS, Leiden University, the Netherlands
knobbe@liacs.nl

² Katholieke Universiteit Leuven, Belgium

³ Eindhoven Technical University, the Netherlands

⁴ Strukton, the Netherlands

⁵ Delft University of Technology, the Netherlands

Abstract. This paper introduces a new project, InfraWatch, that demonstrates the many challenges that a large complex data analysis application has to offer in terms of data capture, management, analysis and reporting. The project is concerned with the intelligent monitoring and analysis of large infrastructural projects in the public domain, such as public roads, highways, tunnels and bridges. As a demonstrator, the project includes the detailed measurement of traffic and weather load on one of the largest highway bridges in the Netherlands. As part of a recent renovation and re-enforcement effort, the bridge has been equipped with a substantial sensor network, which has been producing large amounts of sensor data for more than a year. The bridge is currently equipped with a multitude of vibration and stress sensors, a video camera and weather station. We propose this bridge as a challenging environment for intelligent data analysis research. In this paper we outline the reasons for monitoring infrastructural assets through sensors, the scientific challenges in for example data management and analysis, and we present a visualization tool for the data coming from the bridge. We think that the bridge can serve as a means to promote research and education in intelligent data analysis.

1 Introduction

In practical projects involving data, one often has to model and analyze complex, dynamic systems. An example of this, which is gaining importance, is the monitoring of infrastructural assets such as bridges, tunnels, etc. [1]. Nowadays, the use of advanced sensing and monitoring systems provides the opportunity to collect real-time information from such structures, in order to monitor their performance and to deduce relevant knowledge for decisions on their maintenance demand. Asset owners can use this information to assess the life time perspective of (crucial) infrastructural links and to plan the window within which maintenance can be conducted. When considering the stock of infrastructural assets in

view of service-life assessment, monitoring and sensing systems are very valuable instruments that can be used to extract actual information about its condition and performance.

In terms of condition, sensor systems are mounted in or to structures that monitor the environmental as well as the internal condition. Environmental conditions are related to the climatic changes in which the structure has to be operational, and, in terms of performance, the external and internal actions acting on the structure are recorded. Long and enduring 24/7 monitoring systems are necessary that generate large amounts of data that needs to be evaluated in a smart way such that relevant changes in the data will be noticed and asset management systems will be informed and/or alarmed efficiently.

Managing the huge amounts of data, extracted from the monitoring systems, requires integral knowledge in the field of data management. The best representation for storing the data is not necessarily the best representation for the evaluation of the data. Intelligent Data Analysis is needed for the smart evaluation of data extracted from the degradation mechanisms. The ultimate challenge is to design, develop and optimize a data management system for measuring and reporting the actual performance of large infrastructural projects. Such a system should provide monitoring, notification and reporting services. The goal of such a system is to manage the output of sensors in an optimal way for infrastructure condition assessment.

In this paper we present the InfraWatch project. The goal of this project is to construct a data management system with the above properties for a particular infrastructural asset, the “Hollandse Brug”, one of the Netherlands’ major highway bridges. All the challenges mentioned above are present in this project. We present some initial experiments on the bridge data, and discuss the concrete challenges of this project and argue that it is interesting for the intelligent data analysis community for a variety of reasons. It contains important challenges, but it also provides an attractive and tangible environment for defining data analysis tasks, demonstrating the value of methods, and promoting research and education in intelligent data analysis.

In the next section we introduce the problem of infrastructural asset monitoring and the entailed requirements for data management and analysis on a general level. After that, we will zoom in on the concrete case of the Hollandse Brug.

2 Monitoring infrastructure

In view of asset management of large infrastructural projects, monitoring the actual performance in real-time conditions is becoming indispensable. The actual performance of infrastructural assets in relation to their action loads (traffic, climate, etc.), is considered to be the key element for managing the maintenance requests and for the control of the budgets in the long run. Decisions made regarding the maintenance needs can be considered in view of the technical and economical perspective of an infrastructural asset, and even more importantly, in

view of its functional perspective. Monitoring the performance of infrastructure requires (1) sensor systems to measure, and (2) data management and intelligent data analysis systems for dealing with the large streams of data coming from the sensors. Both are necessary to come up with an optimized system for condition monitoring. In the design of such a system, choices need to be made for

1. The type of sensors: depending on the used materials and other parameters, different types of sensors can be used to trace a structure's actual condition, i.e. sensors for chloride concentration, moisture content, carbonation, etc.
2. The placement of the sensors: the layout and the grid density for the sensors to be positioned in order to get a reliable response of the condition monitoring.
3. The data management: The data received from the sensors will need to be collected, processed and stored by the system and further processed. Results should be communicated by notification and reporting services.
4. Asset Modelling: In addition to the information coming from the data, for prediction of the condition of the asset in the future, modeling of (parts of) the assets is needed. The combination of the data-driven approach with computational modeling gives predictions about the condition of the asset in the future.

Points 1 and 2 are engineering decisions; from the data management and analysis point of view, we assume the type and configuration of sensors are given and we have no control over them. Point 3, data management, contains a number of challenges that hold for infrastructure monitoring in general, and we will discuss these here. The modeling methods relevant for point 4 will strongly depend on the concrete context. Therefore we will discuss point 4 in the next section, after we have introduced the concrete setting of the Hollandse Brug.

2.1 Data Management

Monitoring systems for infrastructural assets are continuously producing large quantities of data. Clearly, it is infeasible to store on location all sensor output at a high resolution, for example more than once per second. Still, for some applications, such detailed data may be required. At the same time, it is unrealistic to perform all computation required for data analysis on-site. Therefore, a sophisticated data management strategy will have to be developed that brings together computing power and the necessary data. This strategy will have to take into account local data collection at different time-resolutions, periodic replication to an off-line data warehouse, scheduled snapshots of intense measurement, as well as some amount of local data analysis for monitoring recent events.

Data collected in monitoring systems roughly serves two purposes:

1. Large scale data mining for analyzing patterns in the stream of sensor output or between different (types of) sensors, as well as analyzing trends over time, for example for finding so-called concept drift.

2. On-line data analysis of (to a certain degree) real-time data for monitoring the integrity of the infrastructure and detecting recent changes.

The first purpose may involve a large range of analysis techniques, and typically requires substantial computational resources. Therefore, this purpose entails the periodic downloading of recent measurement data to a data warehouse, where it can be massaged and analyzed at will. It seems unlikely that all sensor data will be stored at the highest resolution available, at least not for the entire life-span of the infrastructure/monitoring system. The data management system will therefore need to allow for sensor-dependent data storage that may also vary over time. For specific periods of interest, say a notoriously busy day of the year, one may be interested in intense measurement of data, to allow for specific integrated types of data analysis. The effect of varying traffic load on the infrastructure may be assessed by involving video-streams as well as vibration and stress sensors. On the other hand, the influence of weather may be analyzed on a much larger time scale.

The real-time monitoring will clearly have different data management requirements. Such a system component will be using a recent history of sensor output (say one day) to compare these to characteristics of load and infrastructure response over much longer periods, with the intent of detecting recent changes in this relationship. Obviously, this on-line tracking of changes in the infrastructure will have to be done on site, and therefore cannot involve huge amounts of computation. One intended approach is to analyze stress parameters over a longer period off-line, and detect significant patterns and key characteristics that represent nominal behavior of the infrastructure. These patterns may then be uploaded to the monitoring system (with the occasional update), where they can be used to compare to similar results obtained on the recently acquired sensor data. In this way, no large collections of data will need to be stored and processed on-site.

3 The Hollandse Brug

The InfraWatch project is centred around an import highway bridge that is already producing substantial quantities of data: the Hollandse Brug. This is a bridge between the Flevoland and Noord-Holland provinces and is located at the place where the Gooimeer joins the IJmeer (see Figure 1). The bridge was opened in June 1969. National Road A6 uses this bridge. There is also a connection for rail parallel to the highway bridge, as well as a lane for cyclists on the west side of the car bridge. In April 2007 it was announced that measurements would have shown that the bridge did not meet the quality and security requirements. Therefore, the bridge was closed in both directions to freight traffic on April 27, 2007. The repairs were launched in August 2007 and a consortium of companies, Strukton, RWS and Reef has installed a monitoring configuration underneath the first south span of the Hollandse Brug with the main aim to collect data for evaluating how the bridge responds. The sensor network is part of



Fig. 1. Aerial picture of the situation of the Hollandse Brug, which connects the ‘island’ Flevoland to the province Noord-Holland, and the adjacent railway bridge (top).

the strengthening project which was necessary to upgrade the bridges capacity by overlaying.

The monitoring system comprises 145 sensors that measure different aspects of the condition of the bridge, at several locations along the bridge (see Figure 2 for an illustration). The following types of sensors are employed:

- 34 ‘geo-phones’ (vibration sensors) that measure the vertical movement of the bottom of the road-deck as well as the supporting columns.
- 16 strain-gauges embedded in the concrete, measuring horizontal longitudinal stress, and an additional 34 gauges attached to the outside.
- 28 strain-gauges embedded in the concrete, measuring horizontal stress perpendicular to the first 16 strain-gauges, and an additional 13 gauges attached to the outside.
- 10 thermometers embedded in the concrete, and 10 attached on the outside.

Furthermore, there is a weather station, and a video-camera provides a continuous video stream of the actual traffic on the bridge. Additionally, there are also plans to monitor the adjacent railway bridge.

Clearly, the current monitoring set-up is already providing many challenges for data management. For one, the 145 sensors are producing data at rates of 100 Hz, which can amount to a gigabyte of data per day. Adding to that is the continuous stream of video. Although the InfraWatch projects is in its early stages, data is already being gathered and under provisional monitoring. However, the current data available for analysis consists of short snapshots of stress and video data, that is being manually transported from the site to the monitoring location (typically an office environment or Leiden University). One of the aims of the project is to develop sophisticated methods for data management, as outlined in the previous section.

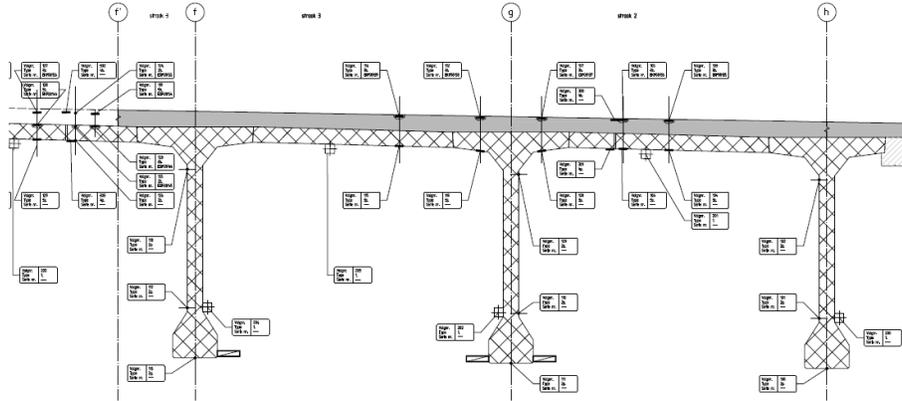


Fig. 2. Detail of the diagram explaining the individual sensor placement

Prior to the start of the InfraWatch project, an initial monitoring application was developed, that allows the visual inspection of both video and sensor information. The application allows the user to navigate through a selected timeframe, and watch the traffic passing over the bridge, while the data over one or more sensors is displayed in synchronised fashion. The user can select the nature of the sensor as well as the location of it, which does not necessarily have to correspond with the location of the camera. Using this application, it is fairly easy to already observe some patterns in the data. For example, the vertical load data nicely corresponds with heavy vehicles passing. However, more sophisticated data analysis should be developed in the course of the project, that also takes into account multivariate behaviour of the data, and spatial relationships between sensors, to name just a few options. In the next sections, we provide some suggestions for the range of analysis approaches this data allows.

4 Data Analysis

We need to distinguish two forms of data analysis: the first form, which we call model construction and which happens offline, consists of analysing data to find patterns in them; the patterns together form a model of the data. The second form, which we call model application, happens online and consists of checking whether the data stream is still consistent with the model. The Hollandse Brug data poses interesting challenges on both sides.

1) Model construction: much data mining research focuses on the model construction task. Many algorithms for detecting patterns in data, and constructing descriptive or predictive models from these patterns, have been described in the literature. The sensor data that we need to deal with here, however, have characteristics that render it impossible to use standard data mining algorithms. First, there is the *temporal* dimension. Each sensor essentially produces a time series of

data. Analysis of time series is a well-investigated problem. However, in this case we cannot analyse each time series on its own: relationships between different time series are relevant. A simple example of this is that a pattern might state that two particular time series normally correlate negatively; but patterns may actually involve much more complicated forms of relationships between (possibly more than two) time series. In addition, these time series may have a different granularity. It is currently not known how such data are best analysed. Second, there is a *spatial* dimension: the sensors are related to each other through their spatial location. The relative position of sensors may be indicated with a graph structure. In that case, patterns may involve combinations of graph structures and time series patterns (for instance, two sensors tend to correlate if they are the same type of sensor and are connected to each other in the spatial graph). It is not obvious how to represent such patterns, and a fortiori no algorithms for discovering them are known. Third, the data are *dynamic*: there may be concept drift, which implies that the patterns relevant at some point in time gradually become less relevant. Models should therefore be adapted regularly. But while a slow shift in the patterns may be normal, a sudden change may indicate a reason for alarm. The question is: how can we distinguish these two different cases?

2) Model application is an equally important task in this context. Model application will happen online, in real-time, with limited computational resources. It is crucial, then, that the developed models can indeed be applied efficiently. This is true for many, but not all models; for instance, for probabilistic graphical models it is known that inference is NP-hard, which makes it non-obvious that they can be applied in this context. The efficient applicability of the learned models is an additional constraint on the data mining task.

Viewed as a whole, we are confronted with data with a complex and evolving relational and spatio-temporal structure. Applying statistical, data mining and pattern recognition techniques to such data is a non-trivial task: there are open questions regarding the optimal representation of the data, how to represent the patterns, what algorithms can be used to detect these patterns in the data (again, existing algorithms will likely not suffice for this task), how to detect significant shifts in the patterns, and how to efficiently detect significant deviations of the data with respect to a given pattern. The development of suitable representations and algorithms to solve these problems is an important research task.

The format of the data and the way it is generated is clearly reminiscent of data streams. The context of this project is somewhat different than what is typically considered in data stream mining: for instance, due to the offline analysis of data, the usual constraints on data stream mining algorithms (namely, that model construction happens online) are less stringent here. This allows us to explore a wider range of algorithms. Nevertheless, it is clear that stream mining is relevant for this project.

In recent years there has been a growing interest in the study and analysis of data streams. Typical examples of such streams include continuous sensor readings. Traditional data mining approaches are not suitable for mining such streams, because they assume static data stored in a database, whereas streams

are continuous, high speed, and unbounded. Therefore, streams must be analyzed as they are produced and high quality, online results need to be guaranteed.

Until now, most pattern mining techniques focus either on non-streaming data, or only consider very simple patterns, such as identifying the hot items from one stream, or constantly maintaining the frequencies in a window sliding over the stream. The challenging task is to extend the existing state-of-the-art into two, orthogonal directions: On the one hand, the mining of *more complex patterns* in streams, such as sequential patterns and evolving graph patterns, and on the other hand, more natural *stream support measures* taking into account the temporal nature of most data streams. Clearly, the classical pattern mining algorithms do not fulfil the constraints imposed on stream processing algorithms. Mining data streams, or stream mining, is therefore a challenging task.

The most popular techniques that have been developed so-far are randomization and approximation, sampling, sketches, and summaries. Randomization and approximation techniques render stream mining algorithms sufficiently fast, at the expense of no longer guaranteeing exactness. Sampling implies that a small sample of the data stream is taken, and costly algorithms are run on the sample. Sketches and summaries help dealing with the abundance of data by instead of storing the complete data stream, which is infeasible, a summary of the relevant features is kept that allows for answering queries about the stream approximately.

5 First experiments

Although the InfraWatch project has only recently started, the sensor network has been up and running for more than a year. During this period, a number of experiments have been performed and specific samples of data have been collected. Some exploratory analysis has been performed to investigate what challenges need to be faced in different aspects of the structural modeling. This section gives some examples.

In theory, one can interpret ‘traffic’ as a series of discrete events, with events being a vehicle passing a particular point at a certain time. However, each individual event will appear to a vibration or load sensor as a signal over some period of time. This temporal spread of the signal is caused by three factors:

1. The physical size of the vehicle. As a vehicle will have a certain length, it will take some time to pass a particular sensor. One can safely assume that this factor is monotone in the length of the vehicle (in the direction of travel) and its speed.
2. The sensitivity area of the sensor. As the sensor is connected to a rigid part of the structure, any movement of the structure will be conducted along it, causing a change in signal of the sensor, even if the vehicle is not exactly located over the sensor. However the effect of the vehicle on the signal will diminish with the distance from the sensor. In effect, the area of sensitivity will act as a form of smoothing on the signal, producing a bump, rather than



Fig. 3. The 10 axle test truck that was driven across the bridge in the early morning.

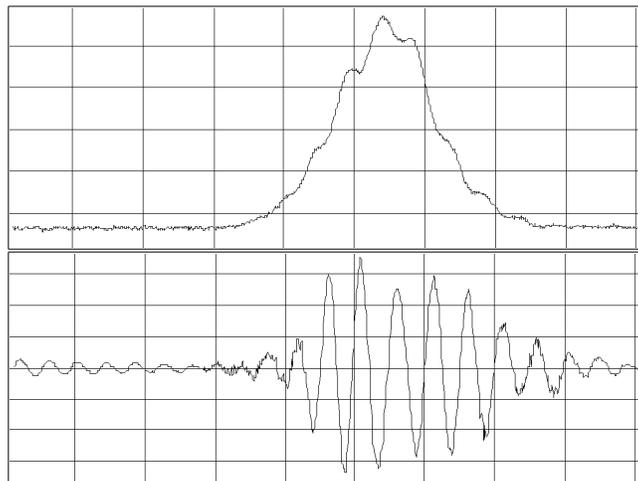


Fig. 4. Measurement of a load (top) and vibration sensor at the moment when the test truck was passing. Individual axles can be observed

a single peak in the sensor data. The effect of this factor will differ between vibration and load sensors, with the latter being bigger, due to complete bridge sections carrying the load of a vehicle.

3. Specific physical properties of the structure, such as the resonance frequency of the bridge. Sudden events, such as a heavy vehicle entering specific sections of the bridge may cause the bridge to subtly sway at a specific frequency that is a physical property of the bridge, and that depends on structural characteristics, such as the size, weight and rigidity of each section. This resonance will cause a signal that starts at the vehicle passing, but that continues for some duration after the event. A Fourier analysis will reveal such dominant frequencies in the spectrum.

One of the essential tasks of the project is to match the continuous signals caused by these three factors with the discrete events of the actual traffic. One way to approach this, is to consider isolated events, and determine their effect

in the sensor-space. Figure 3 shows two pictures of such an isolated test. Trucks were driven with a specific speed (ranging between 50 and 90 km/h) over the sensor network in the early morning, when regular traffic is sparse. Prior to the test, the weight and load distribution over the 10 axles was determined. Different loads were tested, to get a proper variation in examples. Using the resulting data, the sensor-network can effectively be used as a Weigh-In-Motion (WIM) system [2].

Figure 4 shows the effect of a test run on both a load and a vibration sensor. The right graph also shows a subtle vibration of the bridge superimposed on the load signal. This vibration was determined to be approx. 2.5 Hz, over a period of one month. Sudden changes, or gradual drift of this resonance frequency can point to structural degradation of the bridge.

An alternative means of matching continuous signals with discrete events is to remove (or at least minimize) the variable of speed. Figure 5 (left) shows a situation of slow-moving traffic on the far lane of the bridge. By careful manual annotation of consecutive individual video-frames, one can determine the individual events, including some estimate of the size of the vehicle. The right graph shows the effect that the five highlighted vehicles have had on one of the strain-gauges. In such slow-moving conditions, the individual bumps can be identified, and matched to the video-stream. However, there will be a certain amount of ‘stretch’ in the signal, due to the intermittent nature of the passing vehicles. This will make the bumps vary in width in a manner that is somewhat independent of the length of the vehicle.

For the above-mentioned settings, annotation of the video-stream was performed manually, by carefully inspecting individual frames. In order to be able to process large periods of video and sensor-streams, we have been experimenting with automatic detection of vehicles in the images, using a technique for separating the background from the moving traffic (see Figure 6). This technique is flexible and robust, in the sense that it can deal with slight movement of the camera (due to wind and bridge movement), as well as with changing environmental situations (such as weather and lighting). The figure for example shows a rainy day, with a number of large water drops on the lens. Based on the detected location of moving objects, a further aggregation step identifies actual vehicles. The current implementation works fairly consistently, but a clear matching from blobs to events (especially over multiple frames) is still a major challenge.

6 Education Opportunities

Besides being an excellent research challenge and a complex fielded application of Data Mining techniques, the InfraWatch project and its Hollandse Brug are also intended to serve educational purposes [9]. Because of its practical nature, the project will, and has already been an important tool in the teaching of intelligent data analysis techniques to computer science students in the first place. Rather than the traditional focus on basic analysis techniques and algorithms, we now have an opportunity to demonstrate the many complications that tend

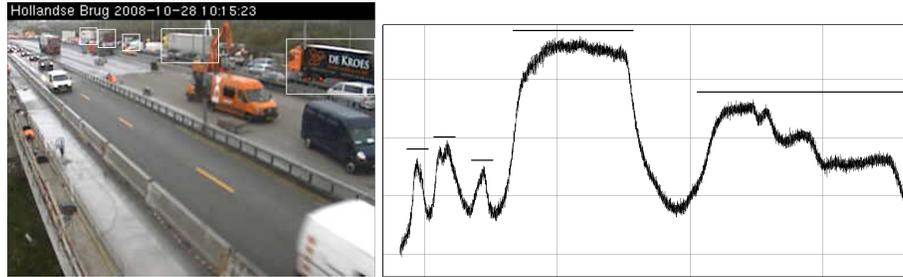


Fig. 5. Slow-moving traffic, and the corresponding output of one of the strain-gauges.



Fig. 6. Estimating large blobs of moving objects: (left) the input image, (middle) the expected background over the recent past, (right) the estimated location of moving objects.

to arise in actual analysis projects [4, 5], and how these should be tackled. These complications include the measuring of data (noise, sensor-failure, ...), the continuous flow of data (data volume, versioning issues, sample rates), the range of analysis paradigms (multivariate analysis, streams, relational aspects), and the inclusion of domain knowledge (spatial aspects, feature extraction). Apart from making the existing data analysis education more attractive and realistic, the project will also serve to attract potential students to analysis-related courses and computer science in general.

7 Conclusion

In this paper we have introduced the InfraWatch project, which has as main goal the setting up of an intelligent infrastructure monitoring system, in particular a data management and analysis system for the Hollandse Brug. It is clear that this system will have online and offline components, and the challenges involved are: determining which functionality is best offered online and offline, determining the optimal representation for online and offline data storage and processing, determining what kind of models are most suitable for this kind of systems, and developing the necessary data analysis techniques for constructing and applying such models.

We believe the project offers a very attractive environment for data analysis for students, scientists and experienced practitioners alike. It provides a tangible and even somewhat spectacular application, with challenges on all levels: students can try to analyse infrastructural data with existing techniques and see what they can find; practitioners can tackle a number of concrete challenges using their expertise on data mining; scientists can study the presented challenges in depth and develop novel techniques and approaches to solve them. Solving the problems defined within the project will require bringing together expertise from very diverse areas in intelligent data analysis, including data and knowledge representation, spatio-temporal data mining, graph mining, sequence mining, data stream mining, computer vision, data visualisation, and more.

Acknowledgements

The InfraWatch project is funded by the Dutch funding agency STW, under project number 10970.

References

1. M. Dejori, H.H. Malik, F. Moerchen, N.C. Tas, and C. Neubauer, 2009 *Development of Data Infrastructure for the Long Term Bridge Performance Program*, In Proceedings of Structures '09, Austin, USA.
2. E. Doupal, R. Calderara, 2004, *Weigh-In-Motion*, In Proceedings of First International Conference on Virtual and Remote Weigh Stations, Orlando.
3. S. Džeroski, H. Blockeel, B. Kompare, S. Kramer, B. Pfahringer, W. van Laer, Experiments in Predicting Biodegradability, In Proceedings ILP 1999, LNCS 1634, 1999 - Springer
4. A. Knobbe, 1997, *Data Mining for Adaptive System Management*, In Proceedings of PAKDD '97, London.
5. A. Knobbe, Bart Marseille, Otto Moerbeek, Daniël M.G. van der Wallen, *Results in Adaptive System Management*, Benelearn'98
6. G. Meijer, *Smart Sensor Systems*, 2008, ISBN: 978-0-470-86691-7, Hardcover, 404 pages.
7. T. Hastie, R. Tibshirani, J. Friedman, 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer Verlag.
8. N. Bessis, 2009, *Grid Technology for Maximizing Collaborative Decision Management and Support: Advancing Effective Virtual Organizations*, University of Bedfordshire, UK
9. R. Gavaldà, 2008, *Machine Learning in Secondary Education?*, In Proceedings TML 2008, Saint Etienne, France, <http://www.lsi.upc.edu/~gavaldà/docencia/tml08-revised.pdf>