ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME

ERCIM
European Research Consortium
for Informatics and Mathematics

# Scientific Report

| | |
|---|---|
| First name / Family name | ALGO CARÈ |
| Nationality | ITALIAN |
| Name of the *Host Organisation* | CWI |
| First Name / family name of the *Scientific Coordinator* | UTE EBERT |
| Period of the fellowship | 01/02/2017 to 31/08/2017 |

## I – SCIENTIFIC ACTIVITY DURING THE FELLOWSHIP

***Research on Machine Learning and System Identification for Space Weather Prediction***

Dr Carè contributed to the "Machine Learning for Space Weather Prediction" project. One of the goals of this project is that of predicting with a good degree of accuracy and reliability the occurrence of extreme magnetic phenomena that could affect human activities. The research focuses on devising and analysing methods that can integrate physical knowledge of the space weather dynamics with information carried by the large amount of data available from solar images, in situ data provided by satellites, and measurements of the magnetic conditions on the earth.

During the fellowship, Dr Carè carried out his research along three main lines.

- **Regression:** Dr Carè has investigated the state-of-the-art in machine learning techniques for Space Weather (in particular, regression techniques based on NARMAX models, Neural Networks and Gaussian Processes), and he wrote an introductory chapter for the book *"Machine Learning Techniques for Space Weather"* on the topic of Regression (see below – Section II, 3 – for the abstract and more information).

- **Classification:** Dr Carè contributed to a paper on the classification of solar wind, see Section II, 4. Discriminating different types of solar wind in an automatic way is preliminary to predicting the effects of the different types of solar wind on the magnetosphere.

- **Uncertainty Quantification:** Assume that a variable of interest depends on a set of uncertain parameters that determine the behaviour of a complex physical system. Evaluating the effects of the uncertainty on the variable of interest is crucial in innumerable situations. A way to address this problem is by modelling the uncertainty by a probability distribution over the uncertain parameters and then computing the corresponding probability distribution of the output. Often, the output probability cannot be computed in explicit form, so that one has to estimate it by resorting to numerical simulations, which, however, are often computationally expensive. A general purpose, adaptive scheme for running simulations in a parsimonious way has been proposed and is currently the subject of investigation on a set of numerical examples.

The space weather project includes members from INRIA, France. Dr Carè was at a group meeting at INRIA on the 21st and the 22nd of February where various possible research lines in Space Weather were discussed, in particular, the possibility of using solar images to improve the forecasting of geomagnetic indices.

***Research on exact, finite-sample identification methods***

The aim of this research activity is devising methods to build mathematical models of unknown systems from data, under minimal assumptions and with precise guarantees on their reliability. In particular, finite-sample identification methods such as SPS (Sign-Perturbed-Sums) and LSCR (Leave-out Sign-dominant Correlation Regions) have been considered.

Previous research in this area has been published or refined in view of publication, in particular

- a journal paper revisiting SPS and LSCR and proposing a new class of methods that gets the best from the previously available approaches has been published (see Section II, 1);

- a conference contribution about the case of unknown model order has been published (Section II, 2);

- previous research on the consistency of SPS for ARX systems has been further refined in view of publication (Section II, 5);

***Research on guaranteed optimization techniques***

In many contexts, a decision has to be made so as to minimise a cost function. When the cost function depends on an uncertain variable, one can make a decision on the basis of some observed realisations of the uncertain variable, which are called scenarios. Once a scenario-based decision has been made, one wants to have some guarantees about the performance of the decision with respect to the future realisations of the uncertain variable. It is a fact that, in some situations of great interest, it is possible to characterise the performance "of tomorrow" without relying on the knowledge of the distribution of the uncertain variable and without using any validation sample. During the fellowship, Dr Carè

- presented a paper on this topic (precisely, on guarantees for least squares optimisation) at the Conference on Computational Management Science, see Section III, 1;

- delivered a one-hour seminar on data-based decisions at the University of Brescia, Italy, on the 2nd of February 2017;

- was invited to give a 90-minute seminar at the Faculty of Mathematics and Physics of the Charles University, Prague (Czech Republic), where every year a cycle of seminars on "Stochastic Programming and Approximation" is held. Dr. Carè's seminar took place on the 30th of March 2017, title: "*Scenario optimization and the risk of empirical costs*".

# II – PUBLICATIONS DURING THE FELLOWSHIP

1. A. Carè, B. Cs. Csáji, M.C. Campi, E. Weyer
   **"Finite-Sample System Identification: An Overview and a New Correlation Method,"**
   IEEE Control Systems Letters ( Volume: 2, Issue: 1, Jan. 2018) online-first [doi: 10.1109/LCSYS.2017.2720969]
   ***Note:*** *The contents of this paper were also selected by CDC 2017 Program Committee for presentation at the 56th IEEE* Conference on Decision and Control*, Melbourne (Australia), 12-15 December 2017.*

2. A. Carè, M.C. Campi, B. Cs. Csáji, E. Weyer
   **"Undermodelling Detection with Sign-Perturbed Sums,"**
   Proceedings of the 20th World Congress of the International Federation of Automatic Control (IFAC WC), Toulouse, France, July 9-14, 2017, pp. 2799–2804

   PAPERS PENDING/ IN PREPARATION:

3. *Status:* submitted for peer review.
   A. Carè,
   **"Regression"** in *Machine Learning Techniques for Space Weather*, E. Camporeale, S. Wing, J. Johnson (Eds.). Elsevier. *(expected publication date: end of year)*
   *Abstract:*
   This chapter presents some well-established ideas and methodologies for the regression problem, which is the problem of learning an input-output relationship from data when the output variable is a real number.
   The contents of this chapter are grouped into four parts, plus a brief introduction to set the terminology.
   In the first part, the regression problem is cast in a probabilistic framework where data are modelled as random variables. In this framework, the concept of prediction error is easily defined and studied, the standard least squares method is introduced, and a discussion on overfitting follows, where some key ideas for balancing complexity and simplicity in choosing a predictor are considered. Finally, the potential of interval predictors is illustrated, and a brief digression on probability density estimation concludes this part.
   In the second part, the regression problem is reformulated as a function approximation problem with no reference to probability, and some basic facts on Neural Networks are presented.
   In the third part, probability is re-introduced in a Bayesian framework, where it is used not only to model the observable data but also to express the observer's uncertainty about the mechanism that generates the data. In particular, some basic facts on Gaussian Process regression are presented.
   In the fourth and last part, input-output relationships where time plays a nonnegligible role are considered, and an introduction to the problem of identifying dynamical systems, both linear and nonlinear, is provided.

4. *Status:* submitted for peer review to the *Journal of Geophysical Research*.
   E. Camporeale, A. Carè, J.E. Borovsky
   **"Classification of Solar Wind with Machine Learning"**
   *Abstract:*
   We present a four-category classification algorithm for the solar wind, based on Gaussian Process. The four categories are the ones previously adopted in the paper "A new four-plasma categorization scheme for the solar wind", by Xu, F., and J. E. Borovsky (published in the Journal of Geophysical Research: Space Physics, 2015): ejecta, coronal hole origin plasma, streamer belt origin plasma, and sector reversal origin plasma. The algorithm is trained and tested on a labeled portion of the OMNI dataset. It uses seven inputs: the solar wind speed, the temperature standard deviation, the sunspot number, the f10.7 index, the Alfven speed, the proton specific entropy and the proton temperature compared to a velocity-dependent expected temperature. The output of the Gaussian Process classifier is a four element vector containing the probabilities that an event (one reading from the hourly-averaged OMNI database) belongs to each category. The probabilistic nature of the prediction allows for a more informative and flexible interpretation of the results, for instance being able to classify events as 'undecided'. The new method has a median accuracy larger than 90% for all categories, even using a small set of data for training. The Receiver Operating Characteristic curve and the reliability diagram also demonstrate the excellent quality of this new method. Finally, we use the algorithm to classify a large portion of the OMNI dataset, and we present for the first time transition probabilities between different solar wind categories. Such probabilities represent the 'climatological' statistics that determine the solar wind baseline.

5. *Status:* draft ready, subject to internal review.
   A. Carè, E.Weyer, B. Cs. Csáji, M.C. Campi,
   **"SPS Confidence Regions for ARX Systems: Exact Guarantees and Asymptotic Properties"**

6. *Status:* accepted for presentation and publication in the *Proceedings of the 56th CDC, Melbourne (Australia), December 2017.*
   F. Baronio, M. Baronio, M.C. Campi, A. Carè, S. Garatti, G. Perone
   **"Ventricular Defibrillation: Classification with G.E.M. and a Roadmap for Future Investigations"**

## III – ATTENDED SEMINARS, WORKSHOPS, CONFERENCES

- **Conference** on Computational Management Science
  Bergamo (Italy), May 30-31, June 1, 2017
  At this conference, organised by the University of Bergamo, the Georgia Institute of Technology and the CMS Journal**,** Dr. Carè delivered a **25 minutes *talk*** about the paper "A Coverage Theory for Least Squares", by A. Carè, S. Garatti and M.C. Campi, which had been previously published on the Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2016.

- Dr Carè plans to attend the **workshop**
  "Space Weather: A Multi-Disciplinary Approach", at Lorentz Center, Leiden (NL). 25 – 29 September 2017.

- Dr Carè attended **seminars** on the topic of *Uncertainty Quantification* at CWI on a regular basis.