Francesco Buccafurri Domenico Saccà (Eds.)

Proceedings of 21st Italian Symposium on Advanced Database Systems SEBD 2013

June 30th – July 04th, 2013

Roccella Jonica, Reggio Calabria, Italy

Francesco Buccafurri Domenico Saccà (Eds.)

Proceedings of 21st Italian Symposium on Advanced Database Systems

SEBD 2013

June 30th – July 04th, 2013
Roccella Jonica, Reggio Calabria, Italy

Volume Editors

Francesco Buccafurri
Dipartimento di Ingegneria dell'Informazione,
delle Infrastrutture e dell'Energia Sostenibile
Università Mediterranea di Reggio Calabria
Via Graziella, Località Feo di Vito
89122 Reggio Calabria, Italy
bucca@unirc.it

Domenico Saccà
Dipartimento di Ingegneria Informatica,
Modellistica, Elettronica e Sistemistica
Università della Calabria
Viale P. Bucci 41C,
87036 Rende (CS), Italy
sacca@unical.it

Multi-level Data Fusion of Environmental Data in Future Internet Applications

Stefano Modafferi, Ajay Chakravarthy, and Zoheir Sabeur

University of Southampton
IT Innovation Centre, Faculty of Physical Sciences and Engineering, United Kingdom
{sm,ajc,zas}@it-innovation.soton.ac.uk

Abstract. The rapid increase in environmental observations which are conducted by SMEs, communities and volunteers using affordable in situ sensors at various scales, together with the more established observatories set up by environmental and space agencies using airborne and space-borne sensing technologies is generating serious amounts of BIG data at ever increasing rates. Furthermore, the emergence of Future Internet technologies and the urgent requirements for the deployment of specific enablers for the delivery of processed environmental knowledge in real-time with advanced situation awareness to citizens has reached greater imminence. It is now highly critical to build and provide services which automate the aggregation of data from various sources, while surmounting the semantic gaps, conflicts and heterogeneity in data sources. The early stages of aggregation of data enable the pre-processing of data generated from multiple sources with the reconciliation between temporal gaps in observation time series, and alignment of their respective asynchronicities. As a result, multi-level processes of fusion need to be implemented and made accessible to large communities of users using future internet services.
This paper presents the process and the preliminary results using RBF networks methods for the spatial fusion of water quality observations and measurements from asynchronous space-borne, in situ and validated models simulation data sources in the Irish Sea.

1 Introduction

In-situ meteorological sensor measurements are generally recorded by sensor hardware at point locations, requiring some form of spatial interpolation if estimates at other locations are needed. Many spatial interpolation methods exist, both deterministic and geostatistical, with accuracies dependent on the nature of the observed phenomena, spatial density of sensors, temporal frequency of sampling and the consistency and accuracy of measurement.

Our case[1] starts considering the Sea Surface temperature in the Irish Sea and, exploiting the phenomena independent algorithms we implement, it is able

[1] The work has been partially supported by the EC under the Envirofi Integrated Project FP7-284898

to move to different dimensions like salinity, chlorophyll and more in general water pollution. These information is relevant for many different business actors and players like fishing boats, touristic cruise organizers and oil companies. Another very important outcome of this applied research concerns the enablement of organisations which are responsible for enforcing the European Water Framework Directives in coastal water basins particularly. They are able to capture levels of water quality parameters in their areas of administrative responsibility. Data fusion and modelling enables the geospatial and temporal merger of fragmented complementary and/or overlapping data sources. For example, in situ and remote sensing data can be set up for fusion and experiment successfully using Radial Basis Function (RBF) networks [1]. The spatial fusion of water quality measurements from satellite and in situ buoys in the Irish Sea have been achieved in the ENVIROFI project [2]. RBF networks are perfect for slow-evolving dimension like sea temperature or salinity and scale well for the spatial data fusion of multiple types of observation sources. They also provide a framework for tuning space scales of advection and diffusion of environmental phenomena at desired and validated levels. Furthermore, the RBF network method generalises well with the increase of data source points since it is a mesh-less technique. The fusion result leads into providing geospatial maps of water quality parameters including areas where observations may be spatially, temporally, or both sparse. Furthermore, uncertainty on these observations may be evaluated for improving operation decision support.

This paper presents the followed approach and some preliminary results achieved by applying it to a practical case in the Irish Sea for fusing sea surface temperature. We outline in section 2 relevant related works, describe the approach in section 4 where we also highlight some preliminary results, while discussing the next steps and concluding in section 5.

2 Related Work

The RBF networks approach [1] is well known and their excellent approximation capabilities have been studied in [6, 7]. In the nineties, solutions of many problems have been based on RBF networks [2, 4, 8]. The recent growing availability of sensor data from Big Data sensor sources has drawn new attention on the RBF technique for spatial fusion where one of its main benefits include the adoption of grid-less computations. These simply involve calculations of Euclidean distances between distributed RBF centres of environmental observations and appropriate RBF projections for spatial data propagation. The approach generalizes and scale to any number of observation centres which do not require to be regularly distributed. In this line, recent works has exploited RBF as a base for risk management and decision support systems [5, 10]. On the other hand merging marine data as base for advanced reasoning has been widely proposed (e.g. see here [11, 3]). The goal of part of the ENVIROFI project and of the research we are presenting in this paper is to provide original contribution merging this

[2] ENVIROFI Project: http://www.envirofi.eu/

research lines in the context of the environment related standard (OGC, SWE, SPS[3])) and with a strong support for the business context and the future internet applications.

3 Fusing environmental data

Fig.1 shows the process we are implementing. From different sources covering different spaces and with different sampling time, we build a unified reconciled model upon which it is possible to provide forecast and extract knowledge.

All the architecture is built on the typical SOA concepts and each module is a service as shown in 2. Support for scalability and plug in of new sources is a key feature that is faced by the transcoding and download module. It is devoted to transcode the data collected from specific data sources to a common format that will be SWE in the future releases while is CSV in the current implementation. If new data source become available, only a specific client for retrieving it has to be built and associated with such module for getting the data into the system. Following a logic flow, the next module is called database feeder. It is in charge of extracting all the metadata (expressed using RDF[4]) that can be directly associated or inferred from the data coming from the sources and storing them in the triplestore. It can optionally store the data in a local database. This is needed when the data sources are not available as continuous service (i.e. the satellite in our case) or for the sake of performances. The use of a hybrid solution (triplestore and classic dbms) for separately storing metadata and data is followed for performance reasons as described in [9].

After the data collections the proper data fusion starts. It follows the JDL methodology and is composed of a low level and high level part. The low level is in charge of pre-processing and semantic reconciling the data in terms of time and in charge of interpolating it. After this phase, a consistent gridless model based on RBF is now available. Usually this part is completed with a data validation procedure. The following step is to use this model for providing spatial and temporal forecasting. A tuned RBF provides such information and it is up to the end user to decide their best use. For help in this final use the framework supports the cross matching of fused data against other data sources for feeding reasoning or alert system (e.g. the additional source can be the scheduling of fishing or guard costal boats).

4 The Marine example

The use case chosen for testing the fusion workflow was in the environmental domain. We chose a region of interest (ROI) along the Irish Sea and initially concentrated on a single phenomenon that is Sea Surface Temperature. Although,

[3] Open Geospatial Consortium: http://www.opengeospatial.org
SPS Sensor Planning Service: http://www.opengeospatial.org/standards/sps
SWE Sensor Web Enablement: http://www.opengeospatial.org/projects/groups/sensorwebdwg
[4] Resource Description Framework: http://www.w3.org/RDF/

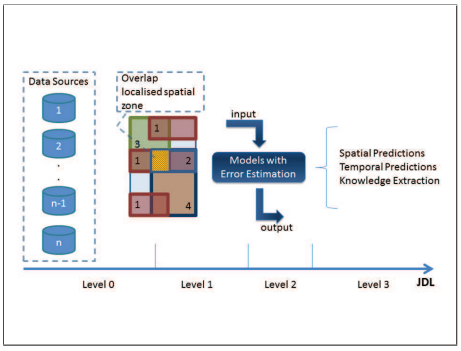

Fig. 1. Overall Process

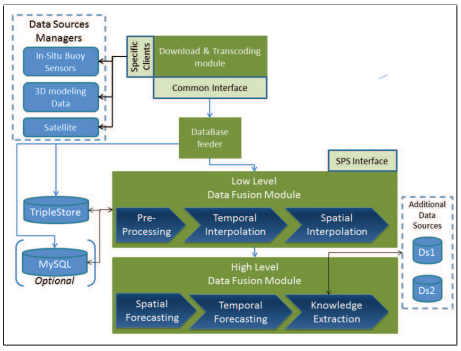

Fig. 2. Architecture

the fusion modules we developed are phenomenon agnostic. The region is defined considering the space covered by the three available data sources as discussed in the next paragraphs. The Irish Marine Institute had deployed various in-situ sensors along the Galway bay. The EUMETSAT geostationary satellite METEOSAT-9 also measured SST over that region. Further, there is a computation weather model deployed in the Marine Institute which predicted SST. This was the ideal use case to perform data fusion from heterogeneous data sources. The high level data fusion is yet under development so it is not presented in the current section. It will address both the parameters forecasting and the knowledge extraction supporting as much as possible scalability and reusability of such module.

4.1 The Data Sources and The Data Sets

The data sources used for the fusion workflow were divided into 3 broad categories. Satellite data sources, in-situ sensor data and 3D computational model data. The geographic area under consideration for the satellite was the coast of Ireland and the surrounding oceanic region with the bounding box coordinates (Upper left corner: 54.50, -12.50 and Lower right corner: 51.00,-6.00). The data was recorded by the METEOSAT-9 geo stationery satellite and hourly recordings of Sea Surface Temperature (SST) values is taken for a period covering from 2011-12-31 to 2013-03-30. The data are available from the EUMETSAT website . Unfortunately the EUMETSAT requires an offline only download process which involved placing orders and the data is available after a few days after the order details have been processed.

The in-situ sensors cover various points in the Region of Interest and measure sea surface temperature. The details of the sensors and their locations are provided in Table1.

The data are available from the Marine Institute's ERDDAP data store as service and the downloading and transcoding module hosts a client for getting them. The third data source, which is also available from ERDDAP is the computational model data for predicting sea surface temperature. The time range

Sensor Name	Location (lon/lat)	Freq.	Time Range
M2	-4.52; 53.48	1 h.	2012-01-01 to 2013-03-30
M3	-10.55; 51.2166	1 h.	2012-06-01 to 2013-03-30
M4	-10; 55	1 h.	2012-01-01 to 2013-03-30
M5	-6.704; 51.69	1 h.	2012-01-01 to 2013-03-30
M6	-15,8813; 53.0748	1 h.	2012-01-01 to 2013-03-30
Galway Bay Wake Buoy	-9.271; 53.277	30 min.	2012-01-01 to 2013-03-30
Belmullet Wave Buoy Berth A	-10.278;54,28	30 min.	2012-01-01 to 2013-03-30
Belmullet Wave Buoy Berth B	-10.146; 54.231	30 min.	2012-01-01 to 2013-03-30

Table 1. In-Situ Sensor Details

chosen for the CONN3D model data is from 2013-12-31 22:00:00 to 2013-03-30 07:00:00. The SST measurements were available at 1 hour intervals. The model uses a ROMS model with a sigma vertical coordinate system which is terrain following (i.e. thickness of levels vary with total water depth within the model domain). In order to measure sea surface temperature a depth of 20 is chosen.

4.2 Low level Data Fusion

The low level data fusion illustrated in the next sections is realized through aggregation and pre-processing and then spatial and temporal interpolation.

Aggregation and Pre-Processing Once the data from heterogeneous data sources (satellite, in-situ and model) are collected, the first step towards achieving data fusion is to perform pre-processing of the datasets such that it is available in a form which could be aggregated into a common schema. Pre-processing includes various sub-processes e.g. format conversion, database import and null value and noise removal. The satellite data is available from GRIB format and covers the entire earth (measuring SST). We had to use the WGRIB2 tool in order to (1) create a sub grid for the region of interest we want to investigate and (2) convert from GRIB2 format to CSV. Once this is done, the database feeder module is used to load the metadata in the triplestore and the data from into a MySQL database. Once the pre-processing step is complete, the data aggregation process is invoked. This also involves various sub processes e.g. unit conversion, date format conversion and schema mapping. The SST readings for satellite data is provided in KELVIN unit while all other data are in Celsius. The format used for recording the timestamp of the measurement is different for satellite (ISO format) and ERDDAP data (UTC). Once the conversion is done, we map the different concepts and terminology into a common schema (in this case it is the Satellite because the most of data come from such source) whose metadata are stored in the triplestore. There were no semantic conflicts found between the various data models. However, mediation between measurement values (e.g unit conversion between temperature values) needed to be done during the pre-processing stage. The size of the aggregated dataset is about 15GB. The

process takes about 22 hours to complete using the database feeder and run in an offline mode overnight.

Spatial-Temporal Interpolation As noted in Table1, the measurement rate for SST for different instruments (satellite, in-situ sensors and model) vary both in time and space. In order to resolve the asynchronicity, both temporal and spatial interpolation need to be performed.

The process of interpolating in-situ sensor and model data is pretty straight forward and we take advantage of several Matlab libraries. We interpolate across time for any given interval (15 min, 20 min etc). The only challenge faced here is the quite big amount of missing values; many of the in-situ measurements had gaps in the data during the considered period. In order to deal with this issue, we use recursive interpolation: i.e. interpolate between the gaps in the data to derive original time range measurements first. However, in the case of satellite data, the resolution of the satellite measurement is very high. The process of interpolating for every given point of the ROI would not be feasible as it would take an enormous amount of time and the amount of data generated would be very huge. In order to address this problem, the ROI is divided into equal grids of a given size and density. The used library takes a vector of lat/long values and a density to create grids covering the area. The more the density the resolution of the grid will be higher.

Next, the mean SST temperature for each grid is calculated and this reading is taken as the input for temporal interpolation rather than considering every single point in the ROI (the assumption here is that given a grid of appropriate size; the SST for a small region does not vary much as SST is a phenomenon which does not vary hugely). The mean SST for a grid would thus provide an accurate picture of the SST values and the process of interpolation would be less time consuming (23 hours, rather than 15 days) and the amount of data produced due to the interpolation process would be manageable (23 GB for a 30 min interval interpolation). Once the process of temporal interpolation is completed, the temporally interpolated dataset for all sources acted as the input to the spatial interpolation process as described in the next section. The time consuming issue lead us to choose a batch approach.

As in the typical use, our RBF has three layers: an input layer, a hidden layer with a non linear RBF activation function and a linear output layer. The temporal interpolation function is what we want to use as activation function. We tried to use a poly-harmonic spline, and then we understood that the classical Gaussian function provides the best results. The use of Matlab libraries provides support in implementing this step in the framework. The interpolation is run on demand and is controllable via the SPS interface. A typical execution lasts less than 1 minute. The following step requires analysing data trend for given sources, evaluating if the trend is regular and make sense and eliminating or studying possible spikes or not justified behaviour. The cleaning process is based on trend visualization tools that help the modeller in evaluating the models. From a practical point, the tuning process requires to "hide" some sources

coming from a known and trusted source, calculating the same from the built model and process and then validating them against the original trusted source. This is a semiautomatic process that ends when the values are enough close one to each other.

Adding a new data source The Fusion architecture is generic and is independent of the data source being used as input into the fusion workflow. The process of adding a new data source is straight forward. Once the data has been acquired from a new source, a configuration file describing the source data e.g. fields like the phenomenon measured, time and data information, location coordinate information etc. need to be specified. The database feeder reads the configuration file and automatically creates the necessary table structures for persistent storage. The metadata information is stored in a semantic triple store and the actual data itself is stored in a relational database. The pre-processing module uses this function of the database feeder to store the data. All the processes henceforth (i.e. aggregation, temporal and spatial interpolation) use the metadata from the triple store to automatically structure their table schemas.

4.3 Preliminary Results

As discussed above the current version uses 3 different data sources and implements the framework up to the low-level data fusion module. Any module is available as service and all the framework is accessible via a standard SPS interface where the end user can specify the desired time and space slice. For any run, a corresponding KML files covering the requested slice is produced. Fig.3 shows an example of the map where the different temperatures are highlighted. The end user can specify his parameter for such process via the SPS interface.

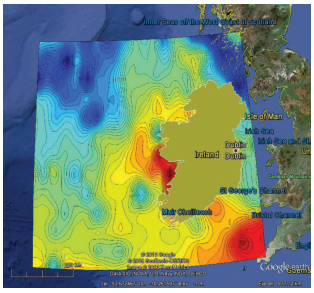

Fig. 3. Spatial Fusion Example over Ireland

5 Conclusion and Future Work

In this paper we have presented an approach for fusing data from various observation sources deployed in the Irish Sea. RBF networks are used in context of the

multi-level JDL data fusion methodology. All the process from the raw data to the knowledge extraction has been structured. The resulting data fusion components provide via a standard SPS interface access to kml files representing spatial fusion of sea surface temperature. Future research work, which exploits the RBF networks scalability, shall involve multiple environmental water quality parameters (e.g. salinity, chlorophyll, dissolved oxygen etc.). With a stable and well performing RBF networks, high level fusion modules such as those specialising in forecast of water quality parameters with dynamic uncertainty shall be implemented. Critical knowledge extraction for decision support becomes possible as a result, particularly for the management of aquaculture and fishing operations in the Irish Sea and beyond. From the environmental regulations point of view as well as compliance with OGC standards, this research work supports the enforcement of the EU Water Framework Directives (Directive 2000/60/EC) and information sharing across European environmental Agencies.

References

1. A.G. Bors. Introduction of the radial basis function (rbf) networks. In *Online Symposium for Electronics Engineers*, volume 1:1, pages 1–7, 2001.
2. V. Chatzis, A. G. Bors, and I. Pitas. Multimodal decision-level fusion for person authentication. *IEEE Trans. on Systems, Man, and Cybernetics, Part A*, 29(6):674–680, 1999.
3. Xiaojun J., Shaohua R., and Yongzhong S. A novel feature-based and application-oriented approach to marine sub-bottom acoustic spatial data fusion. In *Int. Conf. on Geoinformatics (GEOINFORMATICS)*, pages 1–5, 2012.
4. S. Matej and R.M. Lewitt. Practical considerations for 3-d image reconstruction using spherically symmetric volume elements. *Medical Imaging, IEEE Transactions on*, 15(1):68–78, 1996.
5. S. E. Middleton and Z. A. Sabeur. Knowledge-based service architecture for multi-risk environmental decision support applications. In *Proc. of Environmental Software Systems. Frameworks of eEnvironment - 9th IFIP WG 5.11 International Symposium, ISESS*, pages 101–109, 2011.
6. J. Park and J.Sandberg. Universal approximation using radial basis functions network. *Neural Computation*, 3:246–257, 1991.
7. T. Poggio and F. Girosi. Networks for approximation and learning. *Proc. of the IEEE*, 78(9):1481–1497, 1990.
8. R.M. Sanner and J.-J.E. Slotine. Gaussian networks for direct adaptive control. *Neural Networks, IEEE Transactions on*, 3(6):837–863, 1992.
9. TRIDEC_Project. http://www.tridec-online.eu/home.
10. X. Wang and Z. Han. A new disaster monitor and forecast system based on rbf neural networks. In *Proc. of the Int. Conf. on Electrical and Control Engineering*, ICECE, pages 132–136, Washington, DC, USA, 2010. IEEE Computer Society.
11. Weiming X., Gengfeng W., Huirong C., and Xiaodong Y. A fusion method of heterogeneous information for sea-surface wireless sensor networks positioning. In *Computer Distributed Control and Intelligent Environmental Monitoring (CDCIEM), Int. Conf. on*, pages 17–20, 2012.