Smart Water Management for Cities

Klemen Kenda
Jožef Stefan Institute
Ljubljana, Slovenia
klemen.kenda@ijs.si

Stamatia Rizou
Singularlogic
Athens, Greece
srizou@singularlogic.eu

Nikos Mellios
University of Thessaly
Volos, Greece
nmellios@uth.gr

Dimitris Kofinas
University of Thessaly
Volos, Greece
dkofinas@uth.gr

Panagiotis D. Ritsos
Bangor University
Bangor, UK
p.ritsos@bangor.ac.uk

Matej Senožetnik
Jožef Stefan Institute
Ljubljana, Slovenia
matej.senozetnik@ijs.si

Chrysi Laspidou
University of Thessaly
Volos, Greece
laspidou@uth.gr

ABSTRACT

The deployment of real-world water monitoring and analytics tools is still far behind the growing needs of cities, which are facing constant urbanisation and overgrowth of the population. This paper presents a full-stack data-mining infrastructure for smart water management for cities being developed within Water4Cities project. The stack is tested in two use cases - Greek island of Skiathos and Slovenian capital Ljubljana, each facing its own challenges related to groundwater. Bottom layer of the platform provides data gathering and provision infrastructure based on IoT standards. The layer is enriched with a dedicated missing data imputation infrastructure, which supports coherent analysis of long-term impacts of urbanisation and population growth on groundwater reserves. Data-driven approach to groundwater levels analysis, which is important for decision support in flood and groundwater management, has shown promising results and could replace or complement traditional process-driven models. Data visualization capabilities of the platform expose powerful synergies with data mining and contribute significantly to the design of future decision support systems in water management for cities.

CCS CONCEPTS

• **Information systems** → *Data mining*; Specialized information retrieval; • **Computing methodologies** → *Machine learning approaches*;

KEYWORDS

data mining, water management, groundwater, data-driven modeling, machine learning, data visualization

ACM Reference Format:
Klemen Kenda, Stamatia Rizou, Nikos Mellios, Dimitris Kofinas, Panagiotis D. Ritsos, Matej Senožetnik, and Chrysi Laspidou. 2018. Smart Water Management for Cities. In *Proceedings of Fragile Earth workshop at KDD '18 (FEED '18).* ACM, New York, NY, USA, Article 4, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Despite the ongoing research work in the area of smart water management the deployment of advanced high-quality real-time water monitoring tools and services in urban settings is still far from being achieved. Main barriers to the development and adoption of smart water management solutions are related to difficulties in collecting precise monitoring data, lack of interoperability standards and use of simple data mining and data visualization techniques that do not fully exploit the data value. As a response to this challenge, in this paper, we present the Water4Cities approach, which aims to provide novel, beyond state-of-the-art mechanisms for water sensor data collection as well as sophisticated data mining algorithms and data visualization techniques to support two real use cases focusing on water demand management, water reuse and urban planning.

The paper provides an overview of our motivation and the main objectives of the project (Sec.2), the target use cases (Sec.3) and presents preliminary work in the data collection (Sec.4), analysis (Sec.5) and visualization (Sec.6) methods that will be supported by the Water4Cities solution.

2 MOTIVATION AND OBJECTIVES

Urbanisation and overgrowth of population are becoming more intense resulting in overstacked cities, which on the altar of growth and welfare live beyond their environmental capacity and frequently at the expense of vital water resources [13]. Contemporary cities are facing increasing challenges in terms of securing water availability for their citizens, mainly due to insufficient and problematic water management practices and lack of implementing additional measures for climate change adaptation [14]. Moreover, climate change tends to intensify the spatio-temporally uneven distribution of water availability leading to more frequent drought

FEED '18, August 2018, London, United Kingdom

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM.
https://doi.org/10.1145/nnnnnnn.nnnnnnn

or flooding events [4, 15]. Urban water supply often poses a significant and localized stress on proximate water resources, both on quantity and quality.

Groundwater, which constitutes a main provider for urban water supply, is highly impacted, especially in arid or semi-arid areas such as the Mediterranean, by water table degradation on one hand and contamination with various pollutants on the other hand. In coastal areas, degradation of groundwater levels may cause sea intrusion resulting in salinity increase; thus, affecting the suitability of water for a range of uses.

The water saving potential throughout the whole urban water supply chain has been noted as high priority in the field of urban water management. In a different context, the relative impermeability of cities, the inefficient drainage capacity, and the increase of extreme storm events, have made the threat of flooding more intense with negative impact on urban infrastructure, water storage, and eventually on urban life.

The need to improve urban water management efficiency has brought new intrusive approaches to the surface, which are based on sensors, information and communication technologies ICTs), data mining, machine learning, artificially intelligence, decision support systems (DSS) and smart water management; a chain that constitutes the new-age digital water process. Optimized exploitation of water resources, operation of water distribution systems (WDSs), stormwater management, etc, have become core fronts of integrating ICT solutions under a common framework towards sustainability. The integration of smart water management in every facet of urban water activity can reveal the priorities in management issues and facilitate collection of data for raising public awareness and optimizing urban water management.

Smart water management approach intends to secure water resource availability, entailing at the same time water demand effective management, cost effectiveness and environmentally sustainable adaptation [10]. Coupled with proper and thorough planning, ICTs constitute the tool to improve the productivity and efficiency within the water sector, aiming at continuously monitoring water resources towards real-time awareness. Issues such as water security, aging infrastructure, and climate change impact can be addressed or tackled by innovative and smart water management initiatives. The integration of smart water technologies under a common platform will enable water operators and stakeholders to achieve: i) adjusted water management and distribution practices, ii) effective and sustainable economic scheduling, and iii) environmental conservation and protection.

3 USE CASES

The Water4Cities project targets two case studies, each one having a different focus, i.e., one focused around water demand management in the Greek island of Skiathos and another focusing on water reuse and urban planning in the Ljubljana Urban Region. The diversity of the case studies ensures that the Water4Cities solution is not designed for a specific system but can be implemented in a wide range of urban environments, thus making it widely applicable.

3.1 Water Supply in the Island of Skiathos

In the last 4 summers, water shortage crisis events in Skiathos aquifer have occurred during the days of high touristic peak, around the 15th of August. Water demand during these days reaches a high peak following the touristic activity. This peak may rise up at even 3 to 4 times the level of winter water demand. On top of this, touristic activity seems to have an ascending trend, year by year. Water shortage events that have occurred resulted in failing water supply for even a few days. This can have tremendous effects on all aspects of economic activity in the island, on sanitation, public health and well-being. Dealing with such a threat requires having a detailed insight and forecast - short-term and long-term - of the hydraulic and hydrological designing parameters, namely, the aquifer level, the water demand for withdrawals (including non-revenue water), the pressure in the network and all the forecasting parameters such as the weather forecast and the touristic arrivals. This knowledge can allow Skiathos water utility (DEYASK) to apply a tight pressure control scheduling for the days of the crisis, providing for continuous water supply.

DEYASK is currently obliged to assess environmental cost of the water as well as all other cost components and adjust pricing to this assessment, since up to date water is considered to be very underpriced. This kind of assessment requires - except for the hydraulic variables-energy consumption for drilling, transferring, treating, etc., as well as quality data for assessing the quality deterioration. In addition to that, DEYASK will need to have easy access to billing data. This kind of detailed information will provide for building a well justified cost analysis needed for convincing the public for the coming pricing adjustments.

The special interrelation between energy and water (Nexus) is becoming increasingly a hot issue. Investigation of this interrelation seems to reveal hidden amounts of water consumed due to energy consumption or production and hidden amounts of energy consumed due to water consumption or withdrawal. A spatiotemporal analysis of the water-energy nexus and quantification of these amounts may reveal ways to decrease both consumptions, ameliorating the environmental impact through water and carbon footprint reduction. This kind of analysis requires spatio-temporal water and energy consumption and production data.

Skiathos currently faces the issue of groundwater contamination with mercury, while groundwater constitutes the main source of water supply. Except for DEYASK, all involved actors expect to have almost immediate updating regarding mercury levels. Residents, health committee, touristic office, etc., expect to have a thorough insight in water quality status, so that they can be aware of the level of exposure of all users (locals, tourists, health center) to a dangerous pollutant and accordingly regulate their uses in respect to DEYASK instructions. This insight would offer a more integrated overview if it was comparative to quantity variable trends, such as flow-rate, consumption, level of the aquifer table, etc. This comparative presentation of quality and quantity variables would allow stakeholders develop a well justified intuitive link of the two variables, the aquifer level and mercury concentration.

3.2 Groundwater in Ljubljana Aquifer

The Slovenian case study will focus on the Ljubljana Urban Region (LUR), the central region of Slovenia with Ljubljana, the capital of Slovenia. The urban area of the Ljubljana City spreads between two main rivers - the Ljubljanica River and the Sava River, which is the main Slovenian river. The urban and agricultural area between the two rivers is a living and working place for almost 500.000 Slovenian inhabitants. LUR has a long history of various flood protection measures due to its vulnerability to flooding. Despite these measures, many parts of the urban area of the City of Ljubljana are still heavily threatened by the floods, due to intensive urbanization, surface run-off increase as well as climate change effects. In the context of the Slovenian pilot, the Water4Cities will be interconnected with existing GIS systems used for water monitoring and urban planning. In this line, Water4Cities will provide added value capabilities to the current tools, by developing new decision support services for the identification and analysis of groundwater levels, the design of possible Nature Based Solution (NBSs) and related costs and benefits, the maintenance of water infrastructure and potential citizen engagement.

4 DATA GATHERING INFRASTRUCTURE

Water4Cities will use IoT middleware platform described in [12] for almost real-time heterogeneous sources data collection. Apart from collection capabilities our platform provides basic data cleaning, transformation of records into common format, data access protection and usage monitoring. If a data source becomes inactive or inaccessible the system notifies the administrator, so further steps may be taken to resolve the issue.

Data adapters for various data sources are provided, including flat file CSV ingestion, regular scraping of publicly available HTML repositories and ingestion of dedicated REST APIs. It should be noted that Water4Cities will support low power data transmission protocols. In this line, the IoT middleware platform will be integrated (through REST APIs) to an IoT gateway that will collect water related parameters from IoT nodes periodically by employing a low-power wireless radio. In this approach, only the IoT gateway needs to be mains powered, and be internet connected via a 3G/4G connection. Thus, Water4Cities will support low cost water sensors without network connectivity that will be connected to the IoT nodes via appropriate digital interfaces. This setting will be validated in the Greek case study.

As an initial step towards the proof-of-concept validation of the data gathering infrastructure design, we have used the IoT middleware platform to collect three different data sources coming from two case studies: water level and water pump sensor data from Skiathos (Greece), groundwater levels from Slovenia and weather data from both locations. Groundwater data covers whole Slovenia and consists of 518 stations divided into 28 regions. Groundwater level measurements are being collected regularly since 1960, though sampling frequencies are different in different years and some data may be missing due various technical or human-related issues. Records from pump sensors at Skiathos are collected once per day since 2010 and contain information about daily potable water consumption as well as pump working times. Weather data is also collected once per day from various sources.

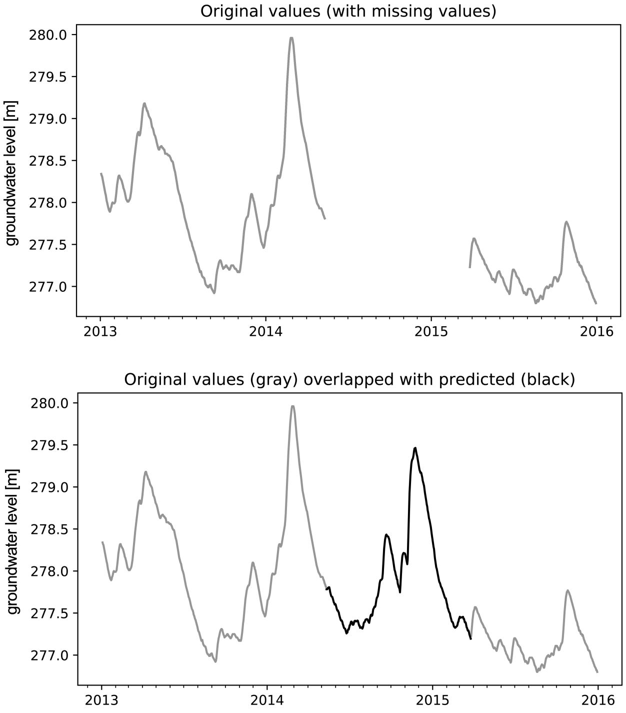

Figure 1: Sub-figure above depicts the missing data for sensor with ID 69 between years 2014 and 2015. The sub-figure below depicts the imputed values, based on our methodology.

All the collected data includes numerical attributes like: information about daily temperatures (minimum, maximum, average), location data, precipitation, sun duration etc. Data are available from 2010 to 2018.

4.1 Missing Data Imputation

As mentioned earlier, one of the main challenges for the smart water management systems is to enable the collection of precise monitoring data. Incorrect or missing data may occur in several scenarios. Consider for instance, the groundwater level monitoring, used at long term to observe changes in the climate and human behaviour can be observed with groundwater dynamics. Long-term availability - usually decades - of the data are needed to address this scenario. During decades sensors break, systems change, groundwater drills move. In this line, often it is difficult if not impossible to find a sensor which would measure data consistently for more than a decade. Additionally, data is often collected with different properties (frequency, precision, etc.). However, intervals of different sensors often overlap, which enables to estimate a missing sensor data with a set of other (available) near-by sensors. In this subsection we present a missing value estimation algorithm, which we tested on data from Ljubljana aquifer, in the context of the preliminary work within Water4Cities project [8].

Table 1: Data imputation results for different (groups of) sensors.

Sensor (group)	R^2	RMSE	Optimal algorithm
highly correlated sensors	0.9976 ± 0.0028	0.0318 ± 0.0244	linear regression
ID 12	0.9388	0.0993	linear regression
ID 73	0.8572	2.7233	random forest

The experiments have been carried out on a subset of Ljubljana city groundwater level sensors[1]. Among 12 selected sensors 10 were highly correlated with others (max $\rho_{X,Y} > 0.98$), and two (with IDs 12 and 73) had lower correlation (max $\rho_{12,Y} < 0.81$, max $\rho_{73,Y} < 0.70$.

The aim of the method is to estimate a missing data in the dataset with available estimators (data from available sensors). We propose construction of multiple models based on different machine learning methods and all available combinations of estimators. Water4Cities data imputation solution implements the following algorithms: linear regression, random forest [2], SVR [5], gradient boosting [6]. At any point in time, based on the available estimators, the optimal model (with smallest error on the test dataset) is chosen.

Results are presented in Table 1. Rough comparison of $\sqrt{R^2}$ and maximal correlation shows significant improvement and results are adequate for additional analysis. Root mean squared error (RMSE) shows that our errors are in the range of centimeters. A real-world example of missing data imputation is presented in Figure 1.

Additional feature engineering (historic values, trends, different aggregates) should better describe underlying processes within the aquifer and is expected to yield even better results. Methodology could also be extended for anomaly detection.

5 DATA-DRIVEN MODELING

Groundwater levels are usually modeled with process-based models. The latter rely on the insightful knowledge of the observed system and processes governing its dynamics. They require many additional spatial data on geological and hydrological properties of the aquifer. Hydrological models usually produce exhaustive spatio-temporal results related to the observed phenomena. When modeling becomes too complex, data driven methods might be beneficial. They require less domain knowledge and they can be built with less diverse data.

One of the goals of data-driven modeling in water management is to predict groundwater levels based on temporal data inputs (historic groundwater and surface water level data, weather and anthropogenic data). The model captures underlying processes based on the data without additional expert user input.

Groundwater level changes through time based on water input and output. It is impossible to predict accurate absolute water level based only on estimates of current water balance. A better problem is to estimate daily groundwater level change.

In the initial stage we have chosen the following inputs: cloud cover, new snow blanket, precipitation, sun duration, different daily temperatures (mean, min, max). We performed extensive feature engineering and analyzed a total of 544 features. New features included different shifted (to determine temporal correlation between weather and groundwater phenomena), averaged and both, shifted and averaged, derivatives of initial input features. Correlation analysis (see correlation matrix in Figure 2) has identified the most relevant features and guided the process of selecting the optimal set of important and uncorrelated features.

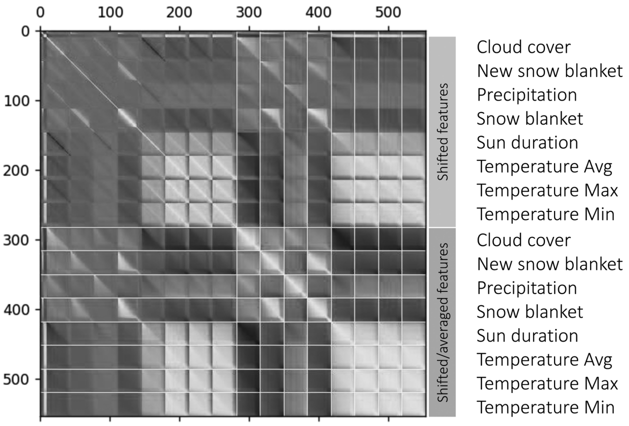

Figure 2: Correlation matrix of 544 input and engineered features. Lighter color represents higher correlation. Each input feature has been transformed (with shifts and/or averages over a moving window up to 100 days). Correlation matrix is divided into 4 similar regions, each depicting correlations among similar attributes (all 4 regions look similar). A significant light block in the bottom right corner shows that sun duration and daily temperatures (average and extremes) are highly correlated. Correlation rises with averaging over multiple days [9].

Table 2: Modeling results of groundwater level change prediction, based on weather data [9].

Method	R^2	RMSE
Linear regression	0.624	2.23×10^{-4} m
Decision trees	0.415	3.46×10^{-4} m
Random forests	0.609	2.31×10^{-4} m
Gradient boosting	0.644	2.11×10^{-4} m

Prediction results on 5 nodes between years 2013 and 2014 are depicted in Figure 3 and Table 2 [9]. RMSE errors show that our error

[1] http://vode.arso.gov.si/hidarhiv/pod_arhiv_tab.php

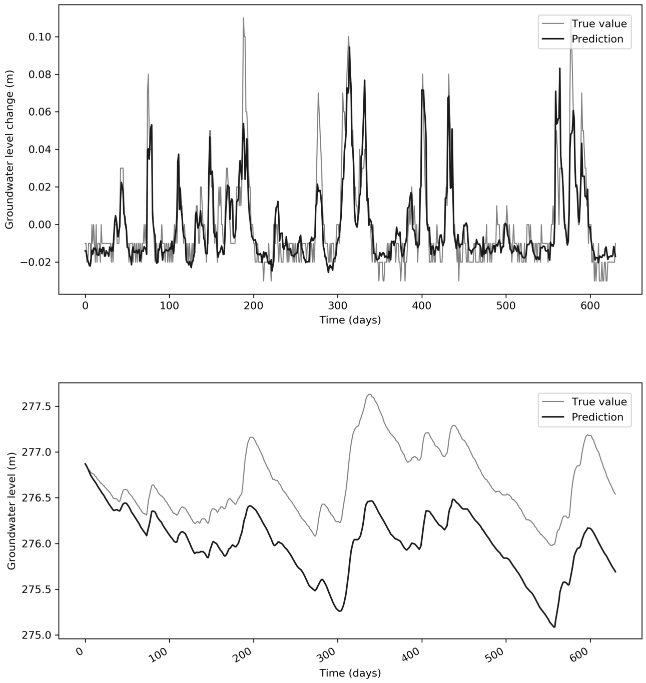

Figure 3: Comparison of true values (gray) and predictions (black) on a groundwater level sensor ID 69 in Ljubljana aquifer. Subfigure above represents comparison of groundwater level changes, the sub-figure below depicts comparison of true groundwater levels vs. predicted cumulative changes.

margin is in the range of 2-4 mm, whereas daily changes of groundwater levels are between -2 to 10cm. R^2 measure is much lower than in missing data imputation, however - the lower sub-figure in Figure 3 which depicts predicted cumulative changes shows that the methodology correctly captures dynamics of the system.

The models can be improved with further feature engineering based on domain-knowledge (i.e. weather data for the whole aquifer - not only the city - would influence water levels, surface water levels are important, etc.).

Additionally, current models are completely independent from groundwater level values. Due to topology of the underground reservoirs this might be a very relevant feature. With precisely defined scenarios (i.e. groundwater level prediction for 3 days in advance) the accuracy could be further improved.

6 DATA VISUALIZATION

One of the most powerful synergies in the expanded data visualization world is the synergy between data mining and information visualization. Researchers of the latter focus on visual mechanisms that provide overview and insight into data distributions [1], whereas researchers of the former use statistical algorithms and machine learning to identify underlying patterns. As claimed by Shneiderman [1], a combinational approach can yield powerful exploration and discovery mechanisms. The synergy between data mining and visualization has been explored for years before the fields of information visualization and visual analytics have formulated into their current incarnations [3, 16]. Mechanisms such as scatter plots, bar charts and histograms, permutation matrices and survey plots have been used in data-mining scenarios, along with techniques such as dimensional brushing, flattening, jittering and global normalization [7].

Water4Cities will employ visualization techniques that facilitate data mining and support the corresponding mechanisms earlier in the Water4Cities pipeline. In particular in terms of water management, a significant portion of the data are spatio-temporal in the form of time-series and event sequences, which will need to be classified and yield tentative predictions, based on historical data. Likewise, the possibility of creating what-if scenarios can assist the decision-making process. These are envisioned to be depicted using small-multiples, for quick overview and inspection. We will also employ visualizations that depict the sensor network’s health and present anomalies such as water consumption spikes/shortages or sensor malfunctions.

In addition to visualizations that depict aggregations or post-processed data, analytical view of the input data can also assist in the decision making (data quality inspection). Especially in the case of externally submitted data, features such as visually detecting missing, erroneous or miss-fielded values (see subsection 4.1), along with type verification and outliers identification can be of great importance to the analytical process.

Currently, we are employing design and ideation methods (e.g. [11]) from the information visualization domain that facilitate the prototyping of data visualization interfaces. Our process emphasizes the user-centered exploration of alternative solutions for our visual interface, ensuring the consideration of diverse data visualization depictions.

7 CONCLUSIONS

In this paper we have presented a full stack solution for data-mining in water management domain. We have described IoT data gathering and provision infrastructure and data-driven approaches towards handling missing data and modeling. The latter has proven to have great potential to complement or even replace traditional process-driven modeling in modern water management decision support systems. We have also opened an issue of using efficient data visualization techniques in combination with data analytics products.

In practice very few stakeholders in the field go beyond SCADA or other IoT systems for monitoring and controlling the water pumps or other devices. Solutions, presented in this paper, bridge the gap to data analytics in water management domain.

There are plenty of further work directions. In data-driven modeling there are a couple of challenges. Firstly, we could improve the methods presented in this paper with additional feature engineering (including additional weather, surface water and anthropogenic data) and bringing additional domain knowledge into the models. Even from the preliminary results, presented in this paper, it can be concluded, that certain dynamics of the system can not be adequately modeled with the current feature set.

Process-driven approaches govern the field of water managements so far. A comparison of data-driven and process-driven approaches is needed in order to determine the actual usability in the field. Data-driven models are much easier to implement and experience from the scientific literature shows, that they are able to provide adequate results. Additionally, previously unknown effects on the water dynamics in an aquifer could be discovered with the help of data-mining approaches.

There is no systematic study of suitability of different machine learning methods in water management domain. Authors often preselect an algorithm with no justification or comparison with others. A study should determine the most suitable machine learning approaches in modeling of a wide variety of water-management datasets.

Deep learning has shown superior results in plenty of the fields. Experience from other fields suggest that often state-of-the-art results can be improved with deep learning; however, improvement might not be significant in comparison to computationally cheaper methods (like random forests or gradient boosting).

Stream mining approaches have been shown to be suitable in energy management. As IoT paradigm is expected to flourish also within the water-management practice in the years to follow it is expected that similar approaches could be used here as well. Streaming methodologies can be implemented in the fields of data cleaning, missing data imputation and predictive analytics.

An extension of the missing data imputation methodology could also be in the field of anomaly detection. Each sensor can be modeled with a set of other sensors, and the results could therefore be used to could cross-validate all the measured values and detect the most obvious outliers.

ACKNOWLEDGMENTS

The work described in this paper has been conducted within Water4Cities project. Project has received funding from the ' Horizon 2020 programme - a Research and Innovation Staff Exchange project under grant agreement number 734409.

The authors would like to thank Filip Koprivec and Matej Čerin for their valuable contribution to the development of data-driven methods presented in this paper.

REFERENCES

[1] 2003. Inventing discovery tools: combining information visualization with data mining. In *The Craft of Information Visualization*, Benjamin B. Bederson and Ben Shneiderman (Eds.). Morgan Kaufmann, San Francisco, 378 – 385. https://doi.org/10.1016/B978-155860915-0/50048-2

[2] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[3] Kristin A. Cook and James J. Thomas. 2005. Illuminating the path: The research and development agenda for visual analytics. (2005).

[4] V Cuculeanu and M Pavelescu. 2013. Climate change and extreme events. *Annals of the Academy of Romanian Scientists* 2, 1 (2013).

[5] Harris Drucker, Christopher JC Burges, Linda Kaufman, Alex J Smola, and Vladimir Vapnik. 1997. Support vector regression machines. In *Advances in neural information processing systems*. 155–161.

[6] Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics* (2001), 1189–1232.

[7] Patrick E. Hoffman and Georges G. Grinstein. 2002. Information Visualization in Data Mining and Knowledge Discovery. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Chapter A Survey of Visualizations for High-dimensional Data Mining, 47–82. http://dl.acm.org/citation.cfm?id=383784.383790

[8] Klemen Kenda, Filip Koprivec, and Dunja Mladenić. 2018. Optimal Missing Value Estimation Algorithm for Groundwater Levels. In *3rd EWaS International Conference*.

[9] Klemen Kenda, Matej Čerin, Mark Bogataj, Matej Senožetnik, Kristina Klemen, Petra Pergar, Chrysi Laspidou, and Dunja Mladenić. 2018. Groundwater Modeling with Machine Learning Techniques: Ljubljana polje AquiferâĂă. In *3rd EWaS International Conference*.

[10] Chrysi Laspidou. 2014. ICT and stakeholder participation for improved urban water management in the cities of the future. *Water Util. J* 8 (2014), 79–85.

[11] Jonathan C. Roberts, Chris Headleand, and Panagiotis D. Ritsos. 2016. Sketching Designs Using the Five Design-Sheet Methodology. *IEEE Transactions on Visualization and Computer Graphics* 22, 1 (2016), 419–428. https://doi.org/10.1109/TVCG.2015.2467271

[12] Matej Senožetnik, Zala Herga, Tine Šubic, Luka Bradeško, Klemen Kenda, Kristina Klemen, Petra Pergar, and Dunja Mladenić. 2018. IoT middleware for water management. In *3rd EWaS International Conference*.

[13] Carlos EM Tucci. 2017. Urbanization and Water Resources. In *Waters of Brazil*. Springer, 89–104.

[14] Hugh Turral, Jacob J Burke, Jean-Marc Faurès, et al. 2011. *Climate change, water and food security*. Food and Agriculture Organization of the United Nations Rome, Italy.

[15] Charles J Vörösmarty, Pamela Green, Joseph Salisbury, and Richard B Lammers. 2000. Global water resources: vulnerability from climate change and population growth. *science* 289, 5477 (2000), 284–288.

[16] Colin Ware. 2012. *Information visualization: perception for design*. Elsevier.