IRIS AperTO

UNIVERSITÀ DEGLI STUDI DI TORINO

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

SIMDMS: Data Management and Analysis to Support Decision Making through Large Simulation Ensembles

This is the author's manuscript

Original Citation:

Availability:

This version is available http://hdl.handle.net/2318/1647179 since 2021-04-29T18:32:50Z

Publisher:

OpenProceedings.org

Published version:

DOI:10.5441/002/edbt.2017.75

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)

07 November 2026

SIMDMS: Data Management and Analysis to Support Decision Making through Large Simulation Ensembles *

Silvestro Poccia Maria Luisa Sapino
University of Torino, Italy
{poccia,mlsapino}@di.unito.it

Sicong Liu Xilun Chen Yash Garg Shengyu Huang
Jung Hyun Kim Xinsheng Li Parth Nagarkar K. Selçuk Candan
Arizona State University, USA
{ s.liu,xilun.chen,ygarg,shuang54,jkim294,lxinshen,pnagarka,candan}@asu.edu

ABSTRACT

Data- and model-driven computer simulations are increasingly critical in many application domains. These simulations may track 100s or 1000s of inter-dependent parameters, spanning multiple layers and spatial-temporal frames, affected by complex dynamic processes operating at different resolutions. Because of the size and complexity of the data and the varying spatial and temporal scales at which the key processes operate, experts often lack the means to analyze results of large simulation ensembles, understand relevant processes, and assess the robustness of conclusions driven from the resulting simulations. Moreover, data and models dynamically evolve over time requiring continuous adaptation of simulation ensembles. The `simDMS` platform aims to address the key challenges underlying the creation and use of large simulation ensembles and enables (a) execution, storage, and indexing of large ensemble simulation data sets and the corresponding models; and (b) search, analysis, and exploration of ensemble simulation data sets to enable ensemble-based decision support.

Keywords

Simulation ensembles, multivariate time series

1. INTRODUCTION

Data- and model-driven computer simulations are increasingly critical in many application domains.

*Research is supported by NSF#1318788 "Data Management for Real-Time Data Driven Epidemic Spread Simulations", NSF#1339835 "E-SDMS: Energy Simulation Data Management System Software", NSF#1610282 "DataStorm: A Data Enabled System for End-to-End Disaster Planning and Response", NSF#1633381 "BIGDATA: Discovering Context-Sensitive Impact in Complex Systems", and "FourCmodeling": EU-H2020 Marie Sklodowska-Curie grant agreement No 690817.

©2017, Copyright is with the authors. Published in Proc. 20th International Conference on Extending Database Technology (EDBT), March 21-24, 2017 - Venice, Italy: ISBN 978-3-89318-073-8, on OpenProceedings.org. Distribution of this paper is permitted under the terms of the Creative Commons license CC-by-nc-nd 4.0

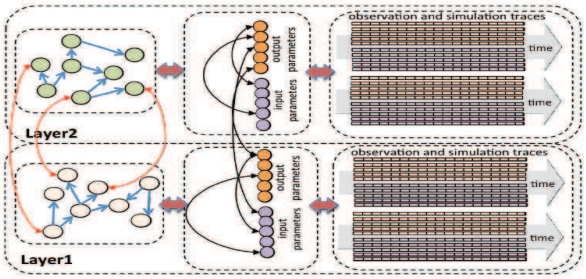


Figure 1: Simulation ensembles are (a) multivariate, (b) multi-modal (temporal, spatial, hierarchical, graphical), (c) multi-layer, (d) multi-resolution, and (e) inter-dependent (i.e., observations of interest depend on and impact each other)

Epidemic Simulation Ensembles: For example, for predicting geo-temporal evolution of epidemics and assessing the impact of interventions, experts often rely on epidemic spread simulation software such as (e.g., GLEaM [2] and STEM [3]). The GLEaM simulation engine, for example, consists of three layers: (a) a population layer, (b) a mobility layer which includes both long-range air travel and short-range commuting patterns between adjacent subpopulations, and (c) an epidemic layer which allows the user to specify parameters (such as reproductive number and seasonality) for the infectious disease, initial outbreak conditions (e.g. seeding of the epidemic and the immunity profile of the subpopulation), and intervention measures.

Building Energy Simulation Ensembles: Similarly, effective building energy management, leading to more sustainable building systems and architectural designs with monitoring, prioritization, and adaptation of building components and subsystems, requires large data-driven simulations involving (a) location and climate information for the city in which the building is located, (b) building construction information, such as building geometry and surface constructions (including exterior walls, interior walls, partitions, floors, ceilings, roofs, windows and doors), (c) building use information, including the lighting and other equipment (e.g. electric, gas, etc.) and the number of people in each area of the building, (d) building thermostatic control information, including the temperature control strategy for each area, (e) heating, ventilation, and air conditioning (HVAC) operation

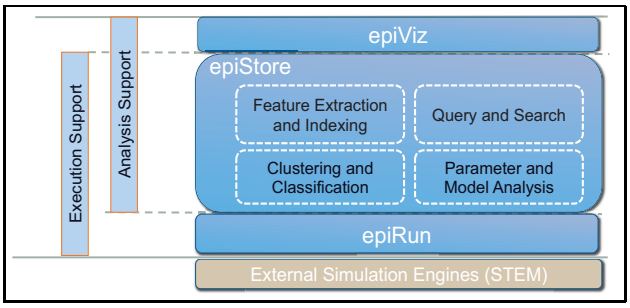


Figure 2: simDMS system overview (instantiated with epidemic simulation ensembles)

and scheduling information, and (f) central plant information for specification and scheduling of boilers, chillers, and other equipment. EnergyPlus software, for example, relies on the description of the building's physical make-up and associated mechanical and other systems and includes time-step based simulation for many energy-related building parameters [1].

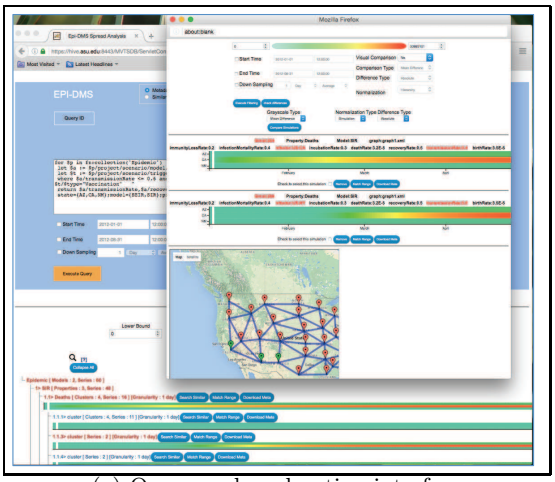
1.1 Challenge: Ensemble based Decisions

While, in most cases, very powerful simulation software exist, using these simulation software for decision making faces several significant challenges: (a) *Creating correct simulation models is a costly operation*, and it is often the case that the designed simulation models are incomplete or imprecise. (b) Also, *the execution of a simulation can be very costly*, given the fact that complex, inter-dependent parameters affected by complex dynamic processes at varying spatial and temporal scales have to be taken into account. (c) *A third major source of cost is the simulation ensemble analysis*: because of the size and complexity of the data and the varying spatial and temporal scales at which the key processes operate, experts often lack the means of analyzing results of large simulation ensembles, understanding relevant processes, and assessing the robustness of conclusions driven from the resulting simulations.

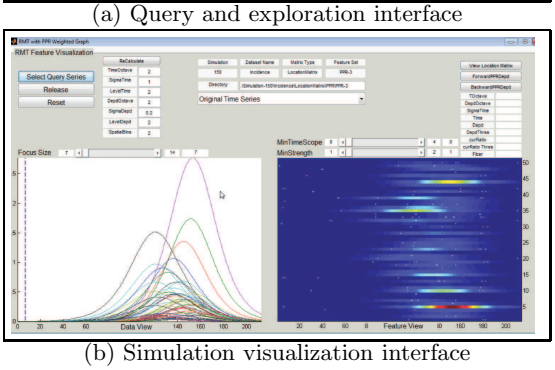
As visualized in Figure 1, the key characteristics of the simulation data sets include the following: (a) multi-variate, (b) multi-modal (temporal, spatial, hierarchical, graphical), (c) multi-layer, (d) multi-resolution, and (e) inter-dependent (i.e., observations of interest depend on and impact each other). In particular, simulations may track 100s or 1000s of inter-dependent parameters, spanning multiple layers and spatial-temporal frames, affected by complex dynamic processes operating at different resolutions. Moreover, generating an appropriate *ensemble* of stochastic realizations may require multiple simulations, each with different parameter settings corresponding to slightly different, but plausible, scenarios. As a consequence, running simulations and interpreting simulation results (along with the real-world observations) to generate timely actionable results are difficult.

We argue that these challenges can be significantly alleviated using a data-driven approach that addresses the following fundamental questions:

- Given a large parameter space and fixed budget of simulations, can we decide which simulations to execute in the ensemble? Can we revise the ensemble as we receive a stream of real world observations?
- Can we compare a large number of simulation ensembles and observations (under different parameter settings) to identify their similarities and differences? Can we analyze one or more simulation ensembles to discover patterns and relationships between input parameters, key events/interventions, and simulation outcomes? Can we discover key events and summarize a large simulation ensemble to highlight these events? Can we classify these key events?
- Can we search and explore simulation ensembles based on the underlying key events or the overall simulation similarities? Can we keep track of the most relevant and most outlier simulations in an ensemble as we receive a stream of real world observations?



(a) Query and exploration interface



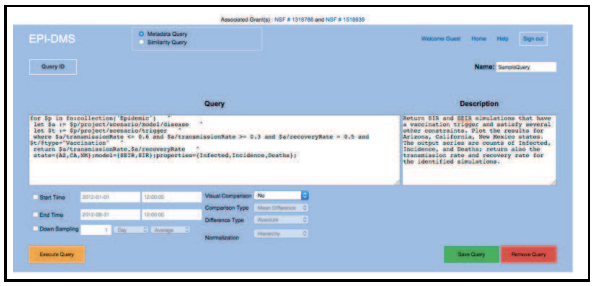
(b) Simulation visualization interface

Figure 3: simVIZ simulation query, visualization, and analysis interfaces (instantiated with epidemic simulation ensembles): visualizing an epidemic simulation as a multi-variate time series and the key robust multivariate (RMT) events [9] identified on a given simulation

1.2 simDMS Overview

The simDMS system (Figure 2) and its visualization engine simVIZ (Figure 3) aim at assisting users to explore large simulation ensembles while limiting the impact of aforementioned challenges [4, 5, 6, 7, 8]. In particular, simDMS supports

- *analysis and indexing of simulation data sets,* including extraction of salient multi-variate temporal features from inter-dependent parameters (spanning multiple layers and spatial-temporal frames, driven by complex dynamic processes operating at different resolutions) and indexing of these features for efficient and accurate search and alignment;




```
FOR $p in fn:collection('EpidemicSimulationEnsemble') ^
LET $diseaseModel := $p/project/scenario/model/disease      ^
LET $triggerModel := $p/project/scenario/trigger      ^
LET $epidemicScenario := $p/project/scenario ^
WHERE
 $diseaseModel/transmissionRate <= 0.6 and
 $diseaseModel/transmissionRate >= 0.3 and
 $diseaseModel/recoveryRate = 0.5 and
 $triggerModel/@type="Vaccination" and
 ($epidemicScenario/infector/@targetISOKey="US-CA"  or
  $epidemicScenario/infector/@targetISOKey="US-NY" ) and
 ($epidemicScenario/graph = "mobility_graph_7.xml"  or
  $epidemicScenario/graph = "mobility_graph_8.xml")  ^
RETURN
 $diseaseModel/transmissionRate,
 $diseaseModel/recoveryRate,
 $epidemicScenario/graph   ^
 STATE={AZ,CA,NM};
 MODEL={SEIR,SIR};
 PROPERTIES={Infected,Incidence,Deaths};
 FROM ={01/01/2012 12:00:00}; TO={08/31/2012 12:00:00};
 BY={1-D}; FUNCTION ={avg};
```

Figure 4: A metadata query over an epidemic simulation ensemble

- *parameter and feature analysis,* including identification of unknown dependencies across the input parameters and output variables spanning the different layers of the observation and simulation data. These, and the processes they imply, can be used for understanding and refining the parameter dependencies and models.

Query and visualization interfaces for the epidemic (epiDMS) and building energy (eDMS) instantiations of the `simDMS` platform can be found at `http://aria.asu.edu/epidms` and `http://aria.asu.edu/edms`, respectively. You can watch a tutorial at `https://youtu.be/9w-4nDhXv3k` .

2. DEMONSTRATION SCENARIOS

We will demonstrate the system on (a) epidemic simulation data sets created using the Spatiotemporal Epidemiological Modeler (STEM) [3] and (b) building energy simulation data sets, created using the EnergyPlus building energy simulation program [1]. The simulations will be stored in `simDMS` and will be visually analyzed during the demonstration using simVIZ.

2.1 Simulation Ensemble Planning

A simulation ensemble (consisting of a set of simulation instances sampled from an input parameter space) can be seen as defining an outcome-surface for each of the output variables (such as the number of deaths that will result from an epidemic): each outcome-surface describes the probability distribution of the potential outcomes for the corresponding variable. These simulation ensembles, consisting of potentially tens of thousands of simulations, are expensive to obtain: therefore we need sampling strategies for the input parameter spaces that eliminate irrelevant scenarios in such a way that more accurate simulation results are obtained where they are more relevant. Moreover, these simulation ensembles need to be continuously revised and refined as the situation on the ground changes: (a) revisions involve incorporating real-world observations into existing simulations to alter their outcomes; (b) refinements involve identifying new simulations to run based on the changing situation on the ground. Therefore, we will demonstrate data-driven sampling strategies to decide (given a budget of simulations) which simulations to run and incremental non-uniform sample-based data construction techniques to revise outcome-surfaces. We will specifically highlight how to assign utility- and cost-functions for each potential sample (based on how well the observed data are fitting the previous simulations, how likely a new simulation at the given sample improves the accuracy of fit, and how costly the corresponding simulation would be) and use these functions to decide the optimal re-sampling strategy.

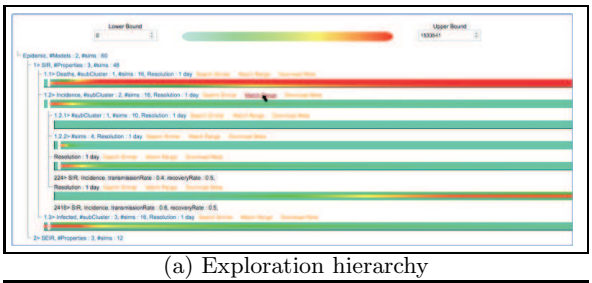
2.2 Scenario- and Similarity-based Querying

A basic function of the `simDMS` system is to retrieve simulations based on a user-specified scenario description. Figure 4 presents a sample query:

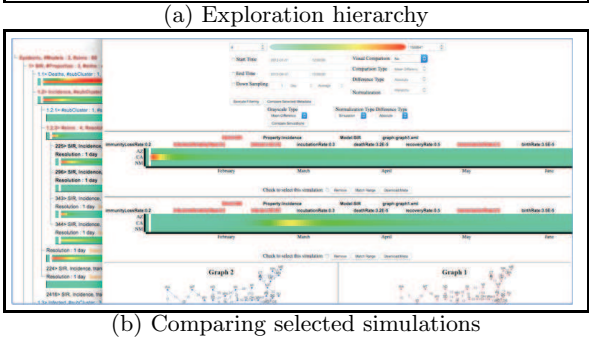
- The "FOR" statement allows the user to select the simulation dataset to query. In this example, the user focuses on the stored simulation set "EpidemicSimulationEnsemble".
- The "LET" statement allows the user to associate variables representing disease and intervention trigger models with epidemic scenarios.
- The "WHERE" clause allows the user to specify conditions on the simulation models to filter those simulations that are relevant for the current analysis. In this example, the user specifies that for the returned simulations, the transmission rate parameter should be between 0.3 and 0.6, the recovery rate parameter should be set to 0.5, and that a "vaccination" type trigger should be included in the simulation model. The user also specifies that epidemic should have started in California (CA) or New York (NY) and the "mobility_graph_7.xml" or "mobility_graph_8.xml" should have been used to generate the simulations.
- The "RETURN" clause lists the simulation parameters to be returned in the result. In this example, the user is interested in the transmission rate, recovery rate, the mobility graph for each returned simulation. In addition, the query asks the system to return the time series corresponding to the "infected", "incidence", and "deaths" simulation output parameters for Arizona (AZ), California (CA), and New Mexico (NM).
- The user further specifies that s/he is interested in only the first 8 months of the simulation.
- Finally, the user specifies that the system returns daily (1-D) averages of the simulation parameters for the specified duration.

Note that, in order to process this single query, `simDMS` combines data of different forms (structured, semi-structured, and temporal), stored in different back-end storage engines.

In addition to scenario-based filtering and search, the platform also enables searching and/or triggering based on particular temporal patterns on the ensembles. This feature

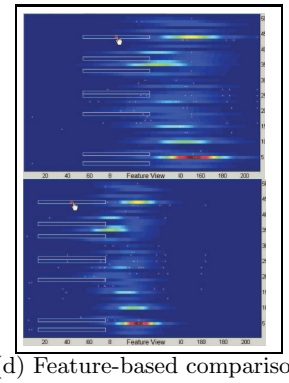


(a) Exploration hierarchy



(b) Comparing selected simulations

(c) Metadata comparison



(d) Feature-based comparison

Figure 5: Sample interfaces for exploring ensembles

allows the expert to identify relevant subsets of stored simulations that match actual real-world observations or specific targets for intervention measures.

2.3 Analysis and Exploration of Ensembles

Once the query is executed and the relevant simulations are identified, the system then organizes the results into a navigable hierarchy, based on the temporal dynamics of the simulation results (Figure 5): Since simulation data sets can be viewed as multi-variate time series, simVIZ focuses on visual analysis (e.g. event detection, similarity and difference analysis) of single and multiple multi-variate simulation data sets. Scenarios that result in similar patterns are grouped under the same branch, while simulations that show major differences in disease development are placed under different branches of the navigation hierarchy. The user can then navigate this hierarchy using "drill-down" and "roll-up" operations and pick sets of simulations to study and compare the corresponding scenarios in further detail.

The interface presents both conventional series plots as well as heatmap visualizations, where each series is shown as a row of pixels. It is important to note that, while the temporal (i.e., horizontal) axis is ordered, the vertical axis corresponding to the different states is not ordered, in that two nearby states according to user mobility may not be neighboring rows on the interface due to the complexity of the mobility graph. The interface also highlights, on the heatmap, the major *robust multi-variate time series (RMT) features* (optimized for supporting alignments of multi-variate time series, leveraging known correlations and dependencies among the variates [9]) identified on the heatmap. An RMT feature is a part of the time series that is *different* in structure from its immediate context in time and/or variate relationships. A key property of these RMT features is that they are *robust* against noise and common transformations, such as temporal shifts or missing variates. This is illustrated in Figure 5(d), which shows two different epidemic simulations, with the same starting state, but different disease parameters and interventions. While the resulting disease evolutions are visibly different in shape, the same multi-variate feature (corresponding to the onset of the disease on the same nearby states) is identified on both simulations. This robustness property of RMT features enables various simVIZ functions, such as search, clustering, classification, and summarization of simulations and large simulation data sets [4, 5, 6, 7, 8].

3. CONCLUSIONS

The `simDMS` platform provides metadata and event-driven analysis and visualization of simulation ensembles to assist decision makers to query and explore ensemble simulations and decide which additional simulations to execute.

4. ACKNOWLEDGMENTS

We also thank Reece Bailey, Hans Behrens, Sarah Fallah-Adl, Ashish Gadarki, Anisha Gupta, Divyanshu Khare, Mao-Lin Li, Nancy Li, Samuel Morton, Shivam Sadakar, Xiaolan Wang, Fiona Zhang, Luke Zhang, and Jerry Zhu for their contributions to the `simDMS` platform.

5. REFERENCES

[1] EnergyPlus Energy Simulation Software. http://apps1.eere.energy.gov/buildings/energyplus/

[2] The Global Epidemic and Mobility Model, GLEAM. http://www.gleamviz.org

[3] STEM. The Spatiotemporal Epidemiological Modeler Project. Available at http://www.eclipse.org/stem.

[4] S. Huang, K.S. Candan, M.L.Sapino. BICP: Block-Incremental CP Decomposition with Update Sensitive Refinement. CIKM 2016: 1221-1230

[5] X. Li, S. Huang, K.S. Candan, M.L. Sapino. 2PCP: Two-phase CP Decomposition for Billion-Scale Dense Tensors. ICDE 2016: pp. 835-846.

[6] S. Liu, Y. Garg, K.S. Candan, M.L. Sapino, G. Chowell. NOTES2: Networks-Of-Traces for Epidemic Spread Simulations. AAAI Workshop on Computational Sustainability, 2015.

[7] S. Liu, S. Poccia, K.S. Candan, G. Chowell, and M.L. Sapino. epiDMS: Data Management and Analytics for Decision-Making From Epidemic Spread Simulation Ensembles. Journal of Infectious Diseases. 2016.

[8] P. Nagarkar, K.S. Candan, A. Bhat. Compressed Spatial Hierarchical Bitmap (cSHB) Indexes for Efficiently Processing Spatial Range Query Workloads. PVLDB 8(12): 1382-1393 (2015)

[9] X.Wang, K.S. Candan, and ML Sapino. Leveraging Metadata for Identifying Local, Robust Multi-Variate Temporal (RMT) Features. ICDE 2014, 388-399.