Dissertation Type: research

DEPARTMENT OF COMPUTER SCIENCE

A Bayesian Inference Engine for UMIS Structured Data

Tom Jager

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Sunday 21st April, 2019

Declaration

This dissertation is submitted to the University of Bristol in accordance with the requirements of the degree of MEng in the Faculty of Engineering. It has not been submitted for any other degree or diploma of any examining body. Except where specifically acknowledged, it is all the work of the Author.

Tom Jager, Sunday 21$^{\text{st}}$ April, 2019

Contents

Executive Summary

A compulsory section, of at most 1 page

This section should précis the project context, aims and objectives, and main contributions (e.g., deliverables) and achievements; the same section may be called an abstract elsewhere. The goal is to ensure the reader is clear about what the topic is, what you have done within this topic, *and* what your view of the outcome is.

The former aspects should be guided by your specification: essentially this section is a (very) short version of what is typically the first chapter. Note that for research-type projects, this **must** include a clear research hypothesis. This will obviously differ significantly for each project, but an example might be as follows:

> My research hypothesis is that a suitable genetic algorithm will yield more accurate results (when applied to the standard ACME data set) than the algorithm proposed by Jones and Smith, while also executing in less time.

The latter aspects should (ideally) be presented as a concise, factual bullet point list. Again the points will differ for each project, but an might be as follows:

- I spent 120 hours collecting material on and learning about the Java garbage-collection sub-system.
- I wrote a total of 5000 lines of source code, comprising a Linux device driver for a robot (in C) and a GUI (in Java) that is used to control it.
- I designed a new algorithm for computing the non-linear mapping from A-space to B-space using a genetic algorithm, see page 17.
- I implemented a version of the algorithm proposed by Jones and Smith in [6], see page 12, corrected a mistake in it, and compared the results with several alternatives.

Supporting Technologies

A compulsory section, of at most 1 page

This section should present a detailed summary, in bullet point form, of any third-party resources (e.g., hardware and software components) used during the project. Use of such resources is always perfectly acceptable: the goal of this section is simply to be clear about how and where they are used, so that a clear assessment of your work can result. The content can focus on the project topic itself (rather, for example, than including "I used LaTeX to prepare my dissertation"); an example is as follows:

- I used the Java `BigInteger` class to support my implementation of RSA.
- I used a parts of the OpenCV computer vision library to capture images from a camera, and for various standard operations (e.g., threshold, edge detection).
- I used an FPGA device supplied by the Department, and altered it to support an open-source UART core obtained from `http://opencores.org/`.
- The web-interface component of my system was implemented by extending the open-source WordPress software available from `http://wordpress.org/`.

Notation and Acronyms

An optional section, of roughly 1 or 2 pages

Any well written document will introduce notation and acronyms before their use, *even if* they are standard in some way: this ensures any reader can understand the resulting self-contained content.

Said introduction can exist within the dissertation itself, wherever that is appropriate. For an acronym, this is typically achieved at the first point of use via “Advanced Encryption Standard (AES)” or similar, noting the capitalisation of relevant letters. However, it can be useful to include an additional, dedicated list at the start of the dissertation; the advantage of doing so is that you cannot mistakenly use an acronym before defining it. A limited example is as follows:

IE	:	Industrial Ecology
IOA	:	Input Output Assessment
LCA	:	Life Cycle Assessment
MFA	:	Material Flow Assessment
UMIS	:	Unified Materials Information System
YSTAFDB	:	Yale Stocks and Flows Database
STAFDB	:	Stocks and Flows Database

Acknowledgements

An optional section, of at most 1 page

It is common practice (although totally optional) to acknowledge any third-party advice, contribution or influence you have found useful during your work. Examples include support from friends or family, the input of your Supervisor and/or Advisor, external organisations or persons who have supplied resources of some kind (e.g., funding, advice or time), and so on.

Chapter 1

Contextual Background

1.1 Industrial Ecology

Industrial ecology (IE) is an area of research focused around the flow of material, energy and money through a system. It can also be described as socio-economic metabolism (SEM) as it models the production, consumption and storage of resources by society. It primarily focuses on the relationship between human originating sectors (anthroposphere) and the natural ecosystem by modelling industrial infrastructures as their own subsystems that interact with their environment [**?**]. The International Society for Industrial Ecology uses the following definition by White, "the study of the flows of materials and energy in industrial and consumer activities, of the effects of these flows on the environment, and of the influences of economic, political, regulatory and social factors on the flow, use and transformation of resources" [**?**]. As such it is a multi-disciplinary field which seeks to account for material and energy data and use it to inform social and economic policy as well as business strategy. The primary motivator for Industrial Ecology is to encourage and ensure sustainable development. By most agreed definitions, this involves ensuring that the economical and societal growth occurs without hampering the ability of development in the future [**?**]. Therefore research in Industrial Ecology focuses on decoupling the relationship industrial development has on natural ecosystems. To do this, the entire life cycle of products and materials are analyzed in order to find areas for greater efficiency and reduced reliance on natural resources. Studies in Industrial Ecology can focus on identifying the amount of flow of specific materials from the anthroposphere into the environment [**?**], to assessing where new stocks of materials are accumulating [**?**], or finding energy and material "loops" which can be closed in order to reuse waste material and energy[**?**].

1.1.1 Industrial Ecology Methodology

There exist a variety of different methodologies to conduct studies in Industrial Ecology. The data resulting from IE studies therefore is usually in a format only suited for that methodology. The three most prevalent methods are Life Cycle Assessments (LCAs), Input-Output Analysis (IOA) and Material Flow Analysis (MFA) [?].

Life Cycle Assessments follow the environmental impact of a product system throughout its life cycle [?]. It often involves compiling an inventory analysis where the life cycle is modelled as a system of processes with material and energy flowing between them. Inputs and outputs from each process are specified with special interest paid to flows into and from the environment. This is used to assess the impact of a product and provide information for corporate decision making in order to improve the efficiency of a product. Other motivators are in reducing a product's environmental impact. LCAs have been known to have sector wide impacts through industry collaboration [?]. Open data formats such as EcoSpold and ILCD as well as shared databases such as Ecoinvent [?], make this possible as different industry partners can combine research and also easily recreate results to ensure accuracy.

Input-Output Analyses takes an economic approach to IE and tracks the flow of money between entities in a geographic region. These can then be used to allocate environmental impact to these entities. As economic data on inter-sector flows are generally more granular than data on the movement of physical material, it can be useful to apply data from IOAs in studies that use other methodologies [?].

Material and Energy Flow Analyses maps the presence of a specific material(s) in a system, paying particular attention to where it accumulates in the form of stocks. It does this by modelling a system as a collection of processes and then accounting for flows between and the build up of stocks in them. This is useful for identifying where cycles can be created and enhanced in these systems in order to increase recycling and therefore reduce a system's dependence on its environment [?]. Static MFAs are where the scope of the analyses falls over a single specified timeframe, whilst dynamic MFAs use data about past and present quantities to estimate future impacts. This can highlight which resources may become scarce in the future and provide warnings about future environmental impacts.

1.2 Unifying Industrial Ecology

Whilst some studies have incorporated data from one methodology into another [?], there is need to provide a common framework for all varieties of industrial ecological data. In [?], Pauliuk et al performed a comprehensive analysis on the myriad of methodologies present in IE. They discovered that each methodology described the system using a shared structure, that of a bipartite directed graph. The shared common property between each system is that material was *transformed* in one process and then *distributed* in a subsequent process. Edges in this bipartite graph denote a flow of material from one a *distribution* node to a *transformation* node or vice versa.

1.2.1 STAFDB and UMIS

Over the past 20 years, the Graedal research group at Yale University have compiled Industrial Ecology data on over 100 materials on a variety of spacial and temporal scales. The data from these studies has been extracted and used to create the Yale Stocks and Flows Database (YSTAFDB) [?]. Ongoing work by Myers, Hoekman and Petard (to be published), is in developing a community driven database for this data called the Stocks and Flows Database (STAFDB). This is an improvement on YSTAFDB as it is designed to be more user friendly and deals with the problem of divergent disaggregation.

Disaggregation is where a property of data (e.g a process that data is coming from or the material being described) is divided into components. In a broad example, data about cars could be disaggregated into electric and non-electric vehicles or large and small cars. If data from two studies with different techniques of disaggregation (or divergent disaggregation) on the same data were structured into the same system, data could be counted twice when performing analysis [?]. To prevent this, process and material data in STAFDB contains a parent field and an `is_separator` flag which serves to keep track of methods of disaggregation and prevent double counting.

Myers et al. has created UMIS, the Unified Materials Information System [?] as a data format to structure data contained in STAFDB. UMIS structures stocks and flows data that comprise IE studies into a format that is agnostic to IOA, MFA and LCA methodologies as a UMIS diagram. This diagram can displayed visually and is also machine readable which allows for greater automation when dealing with stocks and flows data.

UMIS is a step in shifting industrial ecology towards a more virtual platform where industrial ecological data can be stored in a centralized knowledge base, with an ecosystem of tools and routines to develop models and analyze the data for further studies. This would reduce the time taken to perform further studies, allow of greater collaboration in the field and provide greater transparency on the results of studies [?]. An example of such a platform is the Metabolism Of Cities project, a digital research lab created by Paul Hoekman [?]. The platform's primary aim is to encourage research and allow collaboration in studying the metabolism of resources and energy surrounding specific regions. It contains a store of publications, IE data and an online material flow analysis tool to allow users to easily conduct MFA studies and interface with a common online database. Work is currently being done to integrate metabolism of cities with STAFDB to allow for future research to be directly inserted into the database.

UMIS Terminology

The components and design of a UMIS diagram are drawn from MFA concepts but can still be reconciled with IOA and LCA data. A definition list for the components of a UMIS diagram can be found in table ??.

A UMIS diagram can be thought to consist of three layers, a processes and flows layer, a virtual reservoir layer and a metadata layer. The processes and flows layer builds of the work of Pauliuk in [?], and structures the system as flows between transformation and distribution processes in the form of a bipartite directed graph. The virtual reservoir lies on top of this graph and contains information relating to stock. When material is moved in or out of storage in a system, it is represented by moving into or out of the virtual reservoir. As processes represent physical locations in a system, stock are only positioned on top of existing processes. The metadata layer is for storing additional information about stocks flows and processes. This is information concerning the reference space or time frame of a component, source of the data, uncertainty around a quantity's value, unit of the value, calculation details e.t.c. A visualization

UMIS Component	Definition
System Boundary	Definition of the boundary of the data. This is defined by the reference space, time frame and material. Provides a limit of what parts of the real world we are interested in and modelling in this system
Reference Space	The geographical space within which all processes, stocks and flows internal to the UMIS diagram reside.
Reference Material	The material whose stocks and flows are described in the diagram. Can be at a high level of disaggregation (e.g a diagram with a reference material of Steel can contain stocks and flows of Carbon and Iron)
Reference Time frame	The time frame over which material is flowing or is stocked.
Process	An event involving a material
Transformation Process	A process where an object is transformed into another object
Distribution Process	A process where an object is transferred to another process or location
Storage Process	A process where material is moved into or from storage
Stock	Movement of material between a process and storage
Flow	Movement of material between a transformation and distribution process inside the diagram
Cross boundary flow	Movement from material outside the diagram to an internal transformation or distribution process.

Table 1.1: Components of a UMIS diagram

of the key aspects of a UMIS diagram can be seen in figure **??**. In order to ensure that UMIS diagrams are able to be computer generated, transformation processes are enforced to have only one outflow to a distribution process. If a system's design indicates a flow from a transformation process to multiple distribution processes, the processes must be further disaggregated to separate them.

Divergent Disaggregation

UMIS also deals with the divergent dissagregation problem present in processes and materials. This is done by assigning every process and material a parent field and an `is_separator` flag. The parent field contains the process or materials at the next higher level of aggregation. For example the parent field for a blue car material would be car. As a result the processes and materials stored by STAFDB form a tree structure with the `is_separator` flag represent on edges in the tree to detect divergent disaggregation. This thesis focuses on the relationship between flows and processes and therefore considerations about disaggregation are out of scope, but it is likely that the program I have produced will be able to be extended to navigate the aggregation tree and ensure that divergent disaggregation does not occur within a UMIS diagram.

Mass Balancing

MFA, IOA and LCA all share core modelling concepts. Each methodology involve a separation background and foreground systems, with the background system acting as a more generic supplier of inputs

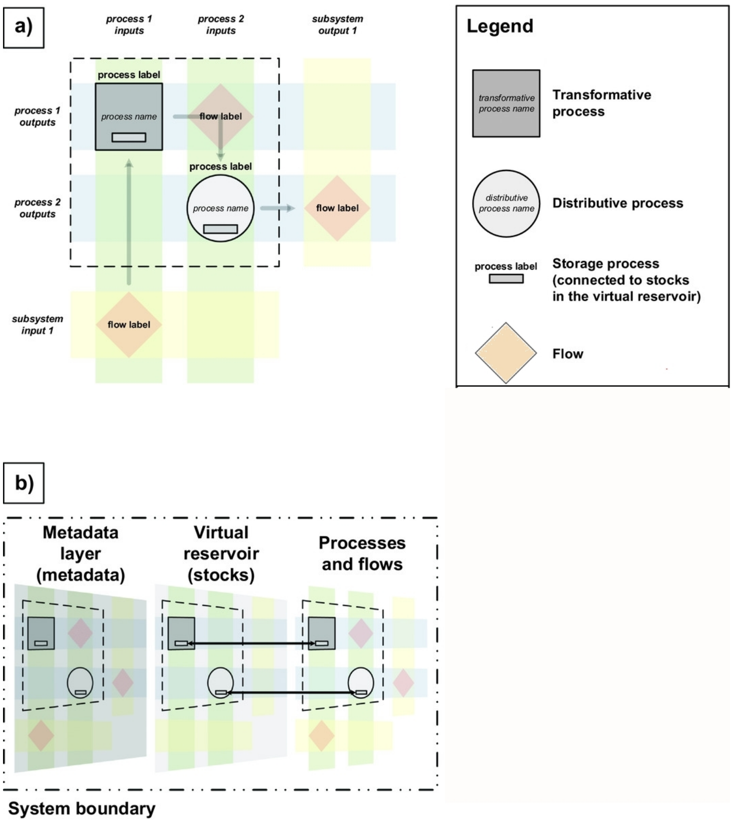

Figure 1.1: (a) Key aspects of a UMIS diagram, visualizing it in a matrix style. Contains transformation, distribution and Storage processes as well as 3 flows. The processes lie on the diagonal and the flows are on the row of the origin process and the column of the destination process (b) The orientation of the virtual reservoir and metadata layer in reference to processes and flows. Flows depicted by grey arrows in (a) and conceptual linkages denoted by black arrows in (b) are for illustration and are not properties of UMIS diagrams. UMIS = Unified Materials Information System. Adapted from [?]

and outputs to the more detailed foreground system [?]. The foreground system is often modelled through a system of equations and constraints, but one common concept to all methodologies is that of mass balancing. The idea of this is that throughout the entire system, matter must be conserved. Therefore the total mass of a given material entering a process must be equal to the total mass leaving [?]. This is defined by the following equation:

Where $i, j \in \{1, .., n\}$ are the processes in the system, f_{ij} is a flow from process i to j, q_i and o_i are the inflow and outflow between process i and the background system respectively, and s_i is the stock being supplied to or coming from storage:

$$\forall i, q_i + \sum_{j}^{n} f_{ji} = \sum_{j}^{n} f_{ij} + o_i \pm s_i \tag{1.1}$$

Further constraints can be supplied in the form of transfer coefficients which allocate the total amount of material flowing into the process amongst the flows leaving the process. With a_{ij} being the transfer coefficient of the flow from process i to j, s_i is stock coming from storage (which may be 0) and z_i as the total of all inflows to process i, this can be defined as:

$$z_i = q_i + \sum_{j}^{n} f_{ji} + s_i \tag{1.2}$$

$$f_{ij} = z_i a_{ij} \tag{1.3}$$

Whilst Myers et al. have demonstrated how material systems in LCA, MFA, IOA studies can be visualized as UMIS diagrams [?], no work has be done to use UMIS structured data in a computational model. In order to demonstrate UMIS's suitability as a structuring system for data in a computational model I will develop a calculation engine for propagating uncertainty through UMIS structured systems.

1.3 Uncertainty in IE

1.3.1 What is uncertainty?

Common practice in IE research is to first define the system in terms of stocks, flows and processes and then identify values for the quantities of stock and flow. These observed values have an associated uncertainty which leads to an uncertainty in model outputs. As IE studies can be used to guide economic and political decisions, it is important to incorporate uncertainty into the results. A study by Danius and Burström [?] found that initial conclusions drawn when studying the flow of nitrogen were inconclusive when data uncertainties were taken into account. Therefore in cases where a comparison of systems are being conducted, the uncertainty around both systems must be taken into account to ensure that the results found are significant and cannot be attributed to variation in underlying data. As such, uncertainty is an important and well studied topic in the fields of LCA, IOA and MFA [?, ?, ?].

The sources of uncertainty can be separated into two kinds [?]. The first is aleatory uncertainty which comes from randomness in the underlying value. This is where either the method of measuring the value has an uncertainty around it so the precise value is unknown, or the value itself is non-deterministic. The second kind is epistemic which refers to generalizations about the value. One cause can originate from when values vary over space and time, therefore the value for a time period (e.g one year) or a geographic area (e.g the entire UK) can have an associated uncertainty. Another source can be subjective judgement such as when an unknown value is estimated based on a known value. Other causes can include scientific disagreement over the true value of a quantity, imprecise language on what the quantity refers to (e.g a material being said to be mostly carbon), and approximation in parameters in order to describe the system as a mathematical model. Epistemic uncertainty is potentially reducible through further investigation, however due to feasibility constraints it is sometimes unavoidable [?].

Further forms of epistemic uncertainty can be caused by possibilities in how the system is defined or described, however as we are looking at a general algorithm for propagating uncertainty in systems, this can be seen as out of scope. Instead, I will focus on uncertainty in specific parameter values.

1.3.2 Incorporating Uncertainty into UMIS

In [?], Laner et al. describe a five step procedure for incorporating uncertainty into MFA studies. These can be found in figure ??:

1. Establish mathematical model
   - Define system elements and relationships between the elements
   - Define equations based on mass balance principle
2. Characterize data uncertainty
   - Evaluate information about data (model parameters, inputs and outputs)
   - Define characterizing functions for uncertainty
3. Combine data and mathematical model
   - Balance model and cross-check data
   - Evaluate plausibility and reconcile data (iterative)
   - Produce a calibrated model using all available data
4. Calculate uncertainty for calibrated model
   - Propagate uncertainty through the model and calculate uncertainty of stocks and flows
   - Interpret uncertainty estimates for resultant values from model
5. Analyze sensitivity & develop scenarios
   - Identify critical model parameters
   - Change parameters to perform scenario analysis

Figure 1.2: Source: [?]

UMIS provides a method for structuring stocks and flows data into the system described in step one of this procedure however does not explicitly specify a unified way for specifying how the uncertainty around data values must be described. Current development on STAFDB and in the Metabolism of Cities project favours an approach investigated by Laner in [?]. This strategy combines the use of a pedigree matrix to classify data quality, as defined by Weidema and Wesnaes [?], and the application of data quality to uncertainty by Hedbrandt and Sörme [?]. The pedigree matrix provides five criteria of data quality (reliability, completeness, temporal correlation and geographical correlation). When a data point is recorded it is given a score for each criteria. The score is combined with a sensitivity level to provide a coefficient of variation. The data point can then be modelled as a normally distributed random variable with a standard deviation related to its coefficient of variation. As stocks and flow values can never be negative and can have asymmetrical properties, it can be appropriate to model the data point as a log-normal random variable. Therefore Laner also provides a method for characterizing data quality as an uncertainty factor which relates to the standard deviation of a log-normal distribution. Therefore Laner's work could be used to convert descriptions of a data point's uncertainty from quality scores to probability density functions.

Whilst the above approach provides a method for characterizing the uncertainty of individual data points independently, it does not provide support for combining the the data points as a system. This is key to steps 3 and 4 of the procedure in figure ??. To do this the data must be arranged as a mathematical model and the uncertainties of each model parameter must be calculated in respect to each other. Furthermore, a framework must be developed to infer uncertainty estimates for values calculated by the model. In [?], Lupton et al. apply a Bayesian inference approach for propagating uncertainty through a model of plastic flows in Austria in 2005. In this paper, Lupton characterizes uncertainty of model parameters as normal distributions. I expand on this approach by generalising it to UMIS specified data and allowing for model parameters to be specified using log-normal distributions also.

1.4 Aims

Below are the aims for this dissertation:

1. Investigate literature in propagating uncertainty through industrial ecology models
2. Develop a program to create a model from UMIS formatted data and propagate uncertainty through it
3. Add support for representing uncertainty through normal, log-normal and uniform distributions
4. Evaluate the accuracy and performance of the program

Chapter 2

Technical Background

A compulsory chapter, of roughly 10 pages

This chapter is intended to describe the technical basis on which execution of the project depends. The goal is to provide a detailed explanation of the specific problem at hand, and existing work that is relevant (e.g., an existing algorithm that you use, alternative solutions proposed, supporting technologies).

Per the same advice in the handbook, note there is a subtly difference from this and a full-blown literature review (or survey). The latter might try to capture and organise (e.g., categorise somehow) *all* related work, potentially offering meta-analysis, whereas here the goal is simple to ensure the dissertation is self-contained. Put another way, after reading this chapter a non-expert reader should have obtained enough background to understand what *you* have done (by reading subsequent sections), then accurately assess your work. You might view an additional goal as giving the reader confidence that you are able to absorb, understand and clearly communicate highly technical material.

Chapter 3

Project Execution

A topic-specific chapter, of roughly 15 pages

This chapter is intended to describe what you did: the goal is to explain the main activity or activities, of any type, which constituted your work during the project. The content is highly topic-specific, but for many projects it will make sense to split the chapter into two sections: one will discuss the design of something (e.g., some hardware or software, or an algorithm, or experiment), including any rationale or decisions made, and the other will discuss how this design was realised via some form of implementation.

Note that it is common to include evidence of "best practice" project management (e.g., use of version control, choice of programming language and so on). Rather than simply a rote list, make sure any such content is useful and/or informative in some way: for example, if there was a decision to be made then explain the trade-offs and implications involved.

Chapter 4

Critical Evaluation

A topic-specific chapter, of roughly 15 **pages**

This chapter is intended to evaluate what you did. The content is highly topic-specific, but for many projects will have flavours of the following:

1. functional testing, including analysis and explanation of failure cases,
2. behavioural testing, often including analysis of any results that draw some form of conclusion wrt. the aims and objectives, and
3. evaluation of options and decisions within the project, and/or a comparison with alternatives.

This chapter often acts to differentiate project quality: even if the work completed is of a high technical quality, critical yet objective evaluation and comparison of the outcomes is crucial. In essence, the reader wants to learn something, so the worst examples amount to simple statements of fact (e.g., “graph X shows the result is Y”); the best examples are analytical and exploratory (e.g., “graph X shows the result is Y, which means Z; this contradicts [1], which may be because I use a different assumption”). As such, both positive *and* negative outcomes are valid *if* presented in a suitable manner.

Chapter 5

Conclusion

A compulsory chapter, of roughly 5 pages

The concluding chapter of a dissertation is often underutilised because it is too often left too close to the deadline: it is important to allocation enough attention. Ideally, the chapter will consist of three parts:

1. (Re)summarise the main contributions and achievements, in essence summing up the content.
2. Clearly state the current project status (e.g., "X is working, Y is not") and evaluate what has been achieved with respect to the initial aims and objectives (e.g., "I completed aim X outlined previously, the evidence for this is within Chapter Y"). There is no problem including aims which were not completed, but it is important to evaluate and/or justify why this is the case.
3. Outline any open problems or future plans. Rather than treat this only as an exercise in what you *could* have done given more time, try to focus on any unexplored options or interesting outcomes (e.g., "my experiment for X gave counter-intuitive results, this could be because Y and would form an interesting area for further study" or "users found feature Z of my software difficult to use, which is obvious in hindsight but not during at design stage; to resolve this, I could clearly apply the technique of Smith [7]").

Appendix A

An Example Appendix

Content which is not central to, but may enhance the dissertation can be included in one or more appendices; examples include, but are not limited to

- lengthy mathematical proofs, numerical or graphical results which are summarised in the main body,
- sample or example calculations, and
- results of user studies or questionnaires.

Note that in line with most research conferences, the marking panel is not obliged to read such appendices.