Leveraging Data Portals as Analytic Platforms: A Case Study of a Simulation of the Queensland Energy and Jobs Plan

William Paul Bell[*] & Nancy Spencer[*]
August 2023

Abstract

This paper discusses how researchers can augment their traditional publications and share their large research datasets in a format easily analysed by other researchers, while enhancing their own analytic capabilities. The case study is the Queensland Energy and Jobs Plan, and the data portal implemented using Power BI. The data portal presents the results from simulating the effect of implementing the plan on the Australian National Electricity Market, encompassing the eastern seaboard of Australia. The case study incorporates 198 simulations which are parameter sweeps of 9 scenarios of coal generation retirement and transmission augmentation, 2 wind levels, and 11 candidate years representing different weather conditions. The portal development process compares different data extraction, transformation, and load strategies and combines proven processes from Business Analysis and Data Management Bodies of Knowledge. The process result is a portal based on a relational database that provides a template for future projects, is simulation model agnostic, informs simulation model improvements, and provides a foundation for more advanced analytics. Future extensions could address dilemmas in the energy market, such as including generation and storage relying on high-frequency arbitrage within long-term development modelling.

Keywords: Power BI; Data Analytics Platform; BABOK; DMBOK, Queensland Energy and Jobs Plan; Agent-based Model of Electricity Systems; Intelligence Energy Systems; Prophet;

JEL Codes: C61, C63, C88, Q40, Q41, Q47, Q48

[*] Centre for Applied Energy Economics & Policy Research, Griffith University.

1 Introduction

Billions of dollars of investment in variable renewable energy (VRE)and storage projects has been and will be committed to enable Australia to reach Australia' zero-net emission (ZNE) target by 2050 (ARENA, 2023). Consequently, many massive project decisions in both the public and private sectors will require support by sound modelling to ensure appropriate levels of investment and returns. While energy generation and storage projects are and can be covered by contracts or spot markets; connecting generation with consumer loads requires transmission and distribution (T&D) networks. Planning and paying for T&D augmentation required for the interconnection of VRE requires coordination, whether centrally planned or via the market modelling will play an essential role.

Too simplistic modelling of power flow requirements is likely to lead to over or underinvestment in transmission and distribution (T&D) capacity. Underinvestment in T&D would lead to generators shedding power or to brownouts and blackouts as consumer demand exceeds the capacity of the T&D. Overinvestment in T&D could lead to underutilised T&D, presenting a loss of opportunity to profitably invest elsewhere.

Additionally, state governments are seeking evidence-based input into the location of new VRE and loads, and appropriate subsidies to incentivise industry to those locations along with local manufacturing content. Financial and simulation modelling can provide this evidence. Furthermore, there is a need for modelling the impacts of different regulations and tariffs to understand the social as well as the economic impacts. Models to achieve these goals are complex and data intensive.

To support Australia's decision-making in the transition to ZNE by 2050 at least cost while maintaining system stability, this paper discusses how researchers can augment their traditional research approaches and share their large research datasets in a format easily analysed by other researchers. Our approach provides the foundation for more advanced analytics to inform agile policy positions.

Section 2 reviews the literature. Section 3 discussed the case study that is a simulation of the Queensland Energy and Jobs Plan (QEJP). Section 4 discusses the motivation and process to normalise the simulation input and output data for use in a relational database within Power BI and publishing to the data portal to the web.[1] Beyond sharing the data for analysis with other researchers, Section 5 discusses nine other benefits from using relational databases within Power BI. Section 6 discusses the costs involved in developing relational data mining and associated databases and analytical tools, while Section 7 discusses the application of relational data mining and analytical tools and identifies possible next steps. Section 8 concludes the paper.

2 Literature review

The novelty in this paper is in synthesising the Business Analysis Bodies of Knowledge (BABoK) (IIBA, 2015) and Data Management Body of Knowledge (DMBoK) (Earley, Henderson, & Data Management Association, 2017) to underpin analytic platform development based on relational databases to support easy reuse of the increasing large datasets from simulation modelling within the energy sector. BABoK has been peer-reviewed

[1] Data Download and Analysis Portal of the Queensland Energy and Jobs Plan 2022 Simulation Modelling Results: https://app.powerbi.com/view?r=eyJrIjoiY2Q2YWVjNzctNDMzMi00NWRmLWEzMzEtZTE2MjdlOTYyNzYxIiwidCI6IjVhN2NjOGFiLWE0ZGMtNGY5Yi1iZjYwLTY2NzE0MDQ5YWQ2MiIsImMiOjEwfQ%3D%3D (The download feature has yet to be enabled).

by thousands of business professionals and DMBoK has been peer-reviewed by thousands of data management and information technology professionals, both are outside the traditional academic literature. Nevertheless, these bodies of knowledge are substantial and rigorously developed. This situation shows a gap in the academic literature.

The benefit of BABoK is the framework it presents to integrate a range of factors within a sector to support decision-making. This benefit is clear in several business sectors, including energy (Takai, Shintani, Andoh, & Washizaki, 2020), records management (Guevara, Loaiza, Lévano, & Zambrano, 2022), education (Sklyar, 2021) and construction (Macariola & Silva, 2019). In all these sectors, the application enables more reliable decision-making based on assimilating substantial amounts of evidence into interpretable chunks.

2.1 Relational Data Mining and the Case Study's Analytics Platform – Power BI

Relational data mining supplies further motivation for building an analytics platform (de Ville, 2001; Džeroski & Lavrač, 2001). Data mining is the practice of analysing large databases to generate novel information using a set of technologies and techniques. An historical note to avoid confusion, relational data mining combined the two disciplines of data mining from statistics and relational databases from data management. Each discipline has its terminology for loosely similar concepts: data mining uses statistical terminology, such as, table, observations, and variables, and relational databases use data management terminology, such as, entity, records, attributes. Traditional data mining methods use a matrix form where rows denote observations, and columns denote variables. This matrix form is unsuitable for analysing the electricity system, given its multiple entities interconnect via information and energy flows. The five main entries include generators, lines, nodes, load serving entities, and independent system operator. A relational database can capture the relationships between these entities and store information about the entities' attributes without redundancy that comes from keeping multiple copies of the same data.

Using the relational database approach efficiently supports the traditional hypothesised impact analysis. However, only using the traditional hypothesis approach will miss relationships hidden in large relational databases that are amenable to relational data mining (Džeroski & Lavrač, 2001) or multi-relational data mining (Padhy & Panigrahi, 2012; Valêncio, Oyama, Neto, & Colombini, 2012). The requirement for the electricity systems to transition to a zero-net emission adds an extra layer of complexity to modelling the electricity system. This requirement makes adopting a relational approach more pressing.

This case study's data analytics platform was developed using Power BI, one of a suite of Microsoft tools. These are part of Microsoft's long-term project to bring together relational databases and data mining along with data extraction, transformation, and loading (ETL) techniques (de Ville, 2001). Microsoft's project matured through various iterations of SQL Server. Currently, Microsoft is expanding its ELT, data mining, and relations database functionality within a new suit of integrated products called Microsoft Fabric[2]. Three of the suite's products, SQL Server, Power BI, and Azure incorporate 'Analysis Services' that include a set of algorithms to compress relational databases called Vertipaq and a data analytics (DAX) language optimised for Vertipaq.[3] Vertipaq is an in-memory columnar

[2] Microsoft Fabric: https://learn.microsoft.com/en-us/fabric/get-started/microsoft-fabric-overview
[3] Microsoft Analysis Services: https://learn.microsoft.com/en-us/analysis-services/analysis-services-overview?view=asallproducts-allversions

database that helps compression and query speed.[4] In a comparative analysis of business intelligence platforms, Gartner finds Power BI in first place for ability to execute and completeness of visions in 2022[5] and 2023[6]. Sections 4 (Step 4) and 5.4 discuss using Vertipaq and DAX in the case study's data analytics platform.

Two proprietary examples of the potential for energy market data analytics platforms include: (1) Intelligence Energy Systems data analytics portal called NEO that is specialised in data visualisation for energy markets[7], and (2) Energy Exemplar's "Energy Analytics and Decision Platform for all Systems" that appears to have integrated its market simulation model PLEXOS with its analytics platform.[8]

2.2 Hybrid Machine Learning and Simulation Modelling and the Case Study's Simulation Model – Agent Base Model of Electrical Systems

New advances in hybrid machine learning and simulation modelling provide even more motivation for building an analytics platform (von Rueden, Mayer, Sifa, Bauckhage, & Garcke, 2020). One of the upcoming issues in modelling optimal development pathways is that some storage and generation is only viable given high-frequency settlement, and modelling optimal development pathways is multi decadal. This situation poses a dilemma. Either make incredibly time-consuming simulations to model high-frequency settlement over future decades with storage remaining viable, or less time-consuming simulations to model lower frequency settlement but lose the viability of storage. In this trade-off, the Integrated Systems Plan (AEMO, 2021, pp. 51-52) opts for either a half-hourly or hourly interval in its high-frequency simulation model to inform its multi decadal capacity model. This half-hourly or hourly modelling is short of the current market settlement period of 5-minutes. The resulting modelling could under-represent viable storage within optimal development pathways given the current settlement period. An alternative approach to this dilemma is applying machine learning to the simulation results (von Rueden et al., 2020). This hybrid machine learning and simulation approach allows leverage of the analytics platform to tackle similar computationally intensive simulations.

The case study's simulation model uses a modified version of Iowa University's open-source model of the US electricity system called Agent-based Model of Electricity Systems (AMES)[9] (Sun & Tesfatsion, 2007). Arguably of the three main groups of simulation models that include equilibrium, dynamics systems (differential equations and discrete event simulations), and complex adaptive systems (agent-based modelling), agent-based models are better suited to modelling the energy transition given the enormous number of entities interacting, several distinct types of entities, and the transition is a non-equilibrium event (Hansen, Liu, & Morrison, 2019; Hoekstra, Steinbuch, & Verbong, 2017; Schimeczek et al., 2023; Shinde & Amelin, 2019; Zhou, Chan, & Chow, 2007).

[4] Vertipaq: https://www.microsoftpressstore.com/articles/article.aspx?p=2449192&seqNum=3
[5] Gartner 2022 review: https://info.microsoft.com/ww-landing-2022-gartner-mq-report-on-bi-and-analytics-platforms.html?LCID=EN-US
[6] Gartner 2023 review: https://www.sisense.com/reports/gartner-magic-quadrant-2023
[7] Intelligent Energy Systems' data portal called NEO: https://iesys.com/NEO/NEO
[8] Energy Exemplar's energy analytics and decision platform: https://www.energyexemplar.com/plexos
[9] Agent-based Model of Electricity Systems (AMES): http://www2.econ.iastate.edu/tesfatsi/AMESMarketHome.htm#AMESVersion

Relevantly, in a comparative analysis of ten different energy market agent-based models, Zhou et al. (2007) find the 'Electricity Market Complex Adaptive System' (EMCAS)[10] the most comprehensive, and AMES lacks AC modelling, only models 4 of 5 EMCAS's ISO functions, and only has basic bidding and planning decision-making. The only Australian model evaluated was CSIRO's NEMSIM (Grozev & Batten, 2006) that focused on carbon emissions, and only modelled the interconnectors between states without intrastate transmission structure. In agreement with Zhou et al. (2007), the CSIRO found that EMCAS was the leading ABM for electricity markets but decided to develop their own given how different the NEM is to the American electricity systems. The EMCAS project seems to have died, given the newest documentation on their website is from 2006, and no response from their contact email address. The AMES project appears ongoing, with version 5 of AMES released July 2020[11] adding built-in support for the fast quadratic programming (QP) optimiser called CPLEX, and some bidding capability that partially addresses one of the earlier criticisms of AMES. However, both EMCAS and AMES were built for the US electricity market and lack any documentation for the NEM. This lack of documentation is discussed further in Section 6.2.

Two electricity market models designed for the NEM are Prophet[12] developed by Intelligent Energy Systems (IES) and PLEXOS developed by Energy Exemplar. Both have extensive documentation for application to the NEM, with sophisticated scenario analysis capabilities. Unlike AMES both use linear programming (LP). LP is faster than QP, but QP can offer more accurate modelling. The Prophet LP model can use either generator linear marginal costs or stack bids. The AMES QP program approximates the stack bids with a quadratic marginal cost function.

There appears a gap for a well-documented QP model of the NEM with sophisticated scenario analysis, and data analytics, to complement the existing LP modelling platforms to develop pathways to zero net emissions. Both Prophet and PLEXOS have considerable investments in datasets and processes supporting their models. Both AMES and Prophet offer free academic licensing, unlike PLEXOS. AMES is no longer free to commercial users since version 5's incorporation of IBM's CPLEX, a QP optimiser, that is only free for academic users. IES supplies a baseline dataset of the NEM for Prophet. Iowa University supplies a several test bed datasets for the US electricity market. The appendix lists other advantages of Prophet over the case study's simulation model. Direct comparisons between LP and QP models using the same input datasets and scenarios could enable confidence intervals and validate simulation results. Presenting the results on the same analytics platform for comparisons could also be beneficial for validating policy positions, given the scale of the investments involved. Confidence and presentation issues are discussed further in Section 7.2.

2.3 Business data modelling versus predictive data modelling

Importantly, the academic energy economics literature and BABoK and DMBoK usage of the term 'model' or 'data model' differs (1) 'predictive data model' that is this paper's case study's simulation model, and (2) 'business data model' that is the data structure of case study's simulation model's input and output data. Business data modelling has two

[10] Electricity Market Complex Adaptive System (EMCAS): https://ceeesa.es.anl.gov/projects/emcas.html
[11] AMES Version 5: http://www2.econ.iastate.edu/tesfatsi/AMESVersionReleaseHistory.htm#AMESV5.00
[12] Intelligent Energy Systems NEM model called Prophet: https://www.iesys.com/Prophet/Index

directions, termed 'forward engineering' and 'reverse engineering'. Each engineering direction has three steps, forward from conceptual to logical to physical, and reverse from physical to logical to conceptual. This paper uses reverse then forward engineering, (1) reverse engineering to develop the conceptual data model from the simulation model's physical data model of its non-relational output, and then (2) forward engineering to develop a new relational physical data model from the newly created conceptual data model. This conceptual data model supplies the abstract ideal form that is hardware and software-agnostic and holds a high-level view of the entities and their attributes. In contrast, the physical data model has the software and hardware implementation details. The intermediate logical step holds details of relationships between entities and attributes but stays hardware and software agnostic. (DMBoK, BABoK). For instance, Guevara et al. (2022) highlights the need to hold the relationships between data details in his work on records management in organisations where it is necessary to keep the links between agendas, minutes, agreements, plans, budgets, progress reports and evaluation reports.

3 Case Study: Simulation modelling of the Queensland Energy and Jobs Plan

The Power BI data portal presents the results from simulating the effect of implementing the Queensland Energy and Jobs Plan (QEJP) on the Australian National Electricity Market (NEM).[13] The QEJP outlines the Queensland Government's pathway to a clean, reliable and affordable energy system to provide power for generations. Key QEJP implementation features include:

- 50% renewable energy by 2030
- 70% renewable energy by 2032
- 80% renewable energy by 2035
- Queensland's Super Grid pathway[14]

The case study models the QEJP's key features as scenarios under differing weather conditions using 198 simulations as follows:

- 9 scenarios (A to I) for the years 2030, 33, and 35 standing for different QEJP transmission augmentations and coal plant closure. See Section xx.
- 2 wind levels (high and low)
- 11 candidate years (2011 to 2022) to stand for different weather conditions and customer demand.
- Product totalling 198 simulations (9 scenarios * 2 wind levels * 11 candidate years.)

Each simulation outputs 42 tables. So, the total number of outputted tables across all simulations is 8,316. Each output table measures the effect of the QEJP on an attribute of the NEM. These 42 attributes grouped by 6 entities, including:

- generators (attributes: emissions, energy generated, energy dispatched)
- transmission lines (attributes: branch flows, branch losses)
- nodes (attributes: spot prices, marginal losses)
- load serving entities (attribute: pumped hydro storage)

[13] Queensland Energy and Jobs Plan: https://www.epw.qld.gov.au/energyandjobsplan/about
[14] Queensland's Super Grid: https://www.epw.qld.gov.au/energyandjobsplan/about/supergrid

- see the data portal's "Simulation Output" tab for details of the outputted 42 tables/attributes, grouped by 6 entities.

The QEJP also includes the requirements for local content in the grid future and three renewable energy zones (REZ). Processes for T&D within the REZ and externally are being formalised and while time limits are articulated, the details are still being finalised.

However, the QEJP only articulates known future generation and transmission investment. It does not capture new energy ventures: their scale, location, or generation characteristics such as ramp speed or recharge. Neither does it capture other new generation and transmission investment in the other eastern seaboard states of Australia. Each newly completed and operational VRE adds further complexity to the model and output data.

To simulate the effect of the QEJP on the NEM, our fellow researchers at CAEEPR used a modified version of Iowa University's open-source model of the US electricity system called Agent-based Model of Electricity Systems (AMES) (Sun & Tesfatsion, 2007). See the Data Portal's page title 'Simulation Model' for a brief description of the modified version of AMES for the QEJP. Four other similar factor impact analysis on the NEM provide more details of the modified AMES model (1) Wind Turbines Generators (Bell, Wild, Foster, & Hewson, 2017; Wild, Bell, Foster, & Hewson, 2015), (2) climate change (Foster et al., 2013), (3) carbon prices (Wild, Bell, & Foster, 2012, 2015) (4), and solar PV (Wild & Bell, 2011).

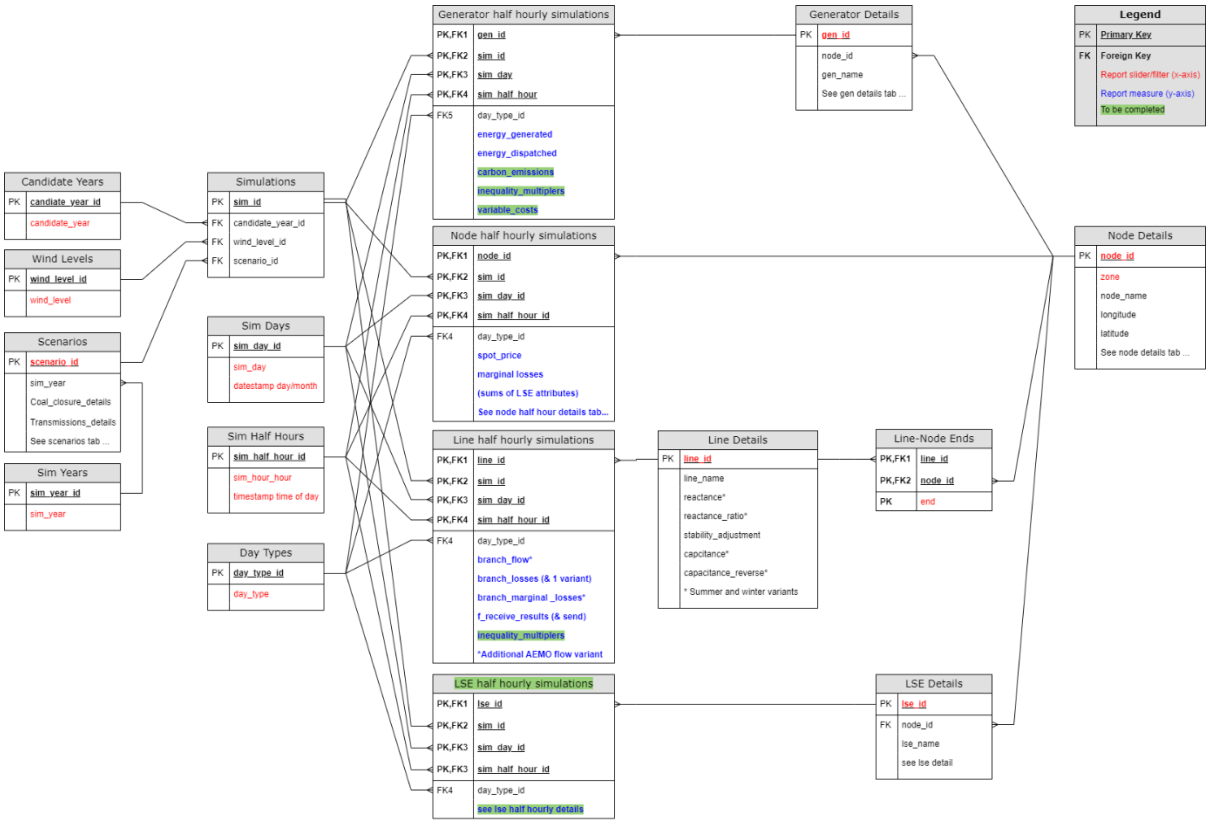
4 Motivation, Method, and Principle-Agent Strategies

This section synthesis data analysis for 'business data models' found in BABoK and DMBoK underpinning the development of an analytics platform based on relational databases.

Motivating the use of techniques from DMBOK and BABOK for the QEJP case study is the requirement to convert the simulation results' 8,316 tabulated text output files distributed among the terminal branches of a three-levelled branching hierarchy directory structure into a form more readily managed and analysed.

The proposed process applied to the case study is stepped out below. These steps outline how to convert the 8,316 output files into a relational database housing all the simulation results in four "normalised" tables, linked together with the details tables as shown in Figure 1. The largest of the four normalised tables are the half-hourly generator, node, and line tables, at respectively 1.2 billion, 167 million, and 244 million rows each. Uploading and linking the normalised tables within Power BI's high performance relational database enables easy analysis and display of the tables.

Figure 1: Entity Relationship Diagram of the case study's output data



To aid clarity, the following steps omit some complications for discussion in the following sections.

1. **Analysing the costs and benefits of alternative strategies, and principle-agent problem**
   In this step, our approach is to analyse the costs and benefits of performing the alternative data ETL strategies: (1) sufficient ETL to answer a single research question or (2) a complete normalisation as shown in Figure 1. The first strategy is the more traditional approach is to perform suffice ETL to complete a single journal paper and the dataset languished with the researcher performing the simulation. This strategy more aligns with the self-interests of the researcher as an agent. In contrast, completing the second strategy as a full normalisation involves a major upfront cost in terms of the researcher's time, with the benefit ensuing from the reuse of the data for multiple research questions and easier access by multiple researchers. The second strategy also aligns more with the principle, in that the organisational benefits outweigh those of the employee of the organisation paying the researcher. Sections 5 and 6 discuss other benefits and costs, respectively, and the principal-agent problem. The principle or full normalisation strategy starts in the next step.

2. **Reverse Engineering to develop an Entity Relationship Diagram**
   In this step, our approach is to develop an Entity Relationship Diagram (ERD) from the outputted tables (BABoK). While this is not difficult, it often takes several diagrams before all the relationships are listed. The conceptual data model or ERD, shown in Figure 1, was developed using the free open-source software called

draw.io. The four main entities are the nodes, generators, transmission lines, and Load Serving Entities (LSE). It is important to correctly determine the ERD before the next steps to prevent rework. Section 5.7 discusses the elimination of two entities, 'NEM' and 'inequality multiplier' found in the output tables.

3. **Cleaning, transforming, and loading the simulation output tables into a relational database.**
   In this step, the grunt work occurs when using SQL databases helps reduce the burden of cleaning and transforming large text output tables.
   a. Clean each of the 198 simulation's 42 tabulated text files while transferring them into a unique database per simulation. The free open-source database software called SQLite[15] was used. The "Lite" refers to the transferability of the database and ease of use, rather than being under powered. For instance, many mobile phone apps use SQLite to store data. SQLite can query the 1.2 billion row generator half-hourly results table and performed throughout the entire process without a problem. SQLite comes with an Open Data Base Connection (ODBC) and other third-party connectors, so can be connected to many development environments. The development environment MATLAB was used in this case study.
   b. Transform the 42 tables within each of the 198 simulation databases into the 4 half-hourly tables as shown in the ERD. This process is known as normalisation and aids computation speed and analysis (see BABOK, DMBOK). The normalisation process in this case involved two main steps. (1) converting "wide" or "pivot" tables into "narrow" or "normalised" tables per attribute. This de-pivoting process involves, converting each column resulting in a set of primary key columns seen in Figure 1 and a single column containing the attribute (2), These normalised tables per attribute merged into an entity table with a single set of primary keys and a single column for each table. For instance, GenOutput table details.
   c. Transform each of the 4 half-hourly tables from each of the 198 simulation databases into 4 half-hourly tables within a single project database.

4. **Forward Engineering to develop a physical data model in Power BI.**
   The next step is to apply forward engineering to implement the conceptual data model/ERD into a physical data model, that happens within Power BI in this case study. Sometimes expediency necessitates that a modified version of the ERD is implemented in the physical model. For instance, if the ERD is exceptional large and complex, the physical model is compartmentalised.
   a. Create detail tables for the four main entities using the simulation input tables.
   b. Create an inversion table between line and node detail tables to eliminate the many-to-many (M:M) relationship between nodes and lines. Relational databases are unable to handle M:M relationships.
   c. Create tables defining the scenarios and simulations.
   d. Create separate tables defining sim_year, sim_half_hour, sim_day. This separation of time into components may seem counterintuitive, but Power BI

[15] SQLite Homepage: https://www.sqlite.org/index.html and DB Browser for SQLite (sqlitebrowser.org)

uses these components as filters and categories to beneficial effect within the relational database discussed in Section 5.

e. Here there are two options depending on user preference. Our investigations have highlighted that Option 1 would be quicker to set up, and Option 2 would provide faster response to queries and data mining because it takes full advantage of Microsoft's Vertipaq relational database compression algorithms and data analytics (DAX) language optimised for Vertipaq. Option 1: Create an ODBC link to the project database and use this ODBC link to connect to Power BI. Option 2: export the individual tables from SQLite to a CSV file and import the CSV files into Power BI.
f. Use Power BI's Model View to link together the tables on their ID fields, as shown in the ERD in Figure 1. The Model View is where the physical data model is designed.
g. Use Power BI's Data View to assign field types, whether numeric or text, and so forth. The Data View is where the details of physical data model are assigned.

5. **Developing graphs, tables, and reports and extending to more complex analytical tools.**
   a. Use Power BI's Report View to create reports that contain graphs and tables from the underlying physical data model. The report view is mostly intuitive.
   b. Section 5 provides more detail on using the data portal's relational database with more sophisticated analytical tools.

6. **Including the input data in the data portal.**
   Ideally, the simulation input data files would be generated from a relational database that would form the initial structure of the data portal's relational database. This approach would extend the relational data mining capability as discussed in Section 5.3 and enable the hybrid machine learning and simulation as discussed in Section 5.9. However, this input file generation from a relational database approach was not considered in the initial version of the QEJP simulation project. The result was massive data redundancy in the 198 input files, each holding over 40 tables. This massive redundancy would increase the time to complete an ETL well beyond any cost benefit for the researcher completing the ETL. Section 5.7 discusses other simulation model improvement insights gained from applying 'business data modelling' to the simulation model input and output data for the next project.

5 Leveraging the benefits from developing the portal

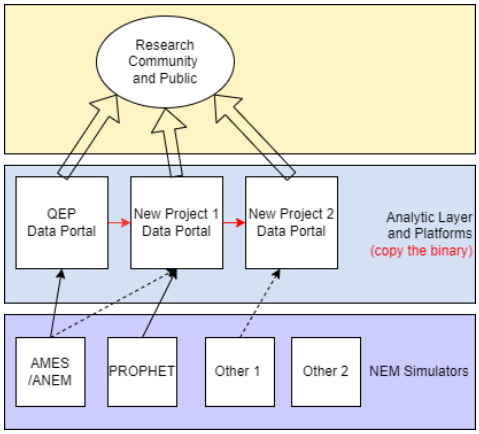
The six steps in Section 4 describe the first stage in developing a portal that enables basic reporting with graphs and tables. This section describes how to leverage the above investment in ETL of the simulation output results tables into a relational database within a data analytics platform.

5.1 The template effect and portals as an analytic layer between simulation models and researchers

The data portal separates the analytics layer from the simulation layer and illustrated in Figure 2. This separation has at least three benefits: (1) The ability to build up the analytic capability through iterative copying and reuse of the platform for each new project; (2) The separation allows an agnostic approach to selecting a simulation model, where a simulation

model can be swapped between projects or the results of two or more simulation models compared to help confirm conclusions or supply confidence intervals; and (3) The users of the data portal have a familiar interface between projects, so can reduce learning time.

Figure 2: Data portals as platforms providing a common analytical interface between simulation results and the research community and undergoing iterative improvements between projects.

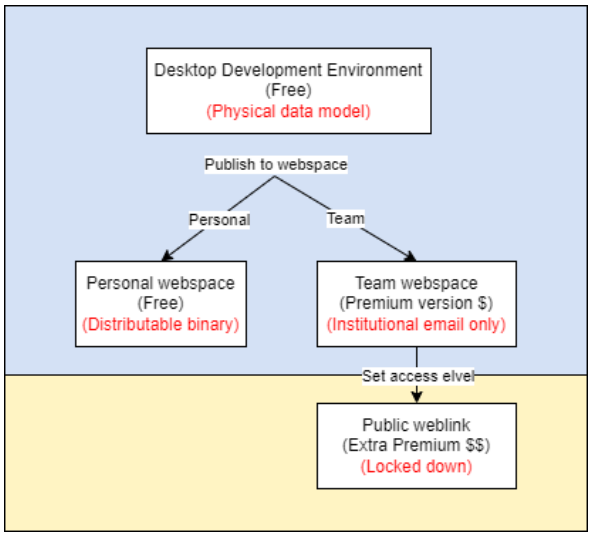


5.2 Using the portal's data compression to share large, structured datasets.

Figure 3 shows the Power BI development environment, publication pathways, and costs. The size of the case study's portal in desktop development environment is 5 GB. This binary file compresses the more than 100 GB of imported CSV files. This small size readily allows the template effect discussed above and provides the opportunity to share the portal development environment with other researchers.

In addition to the developer's own free personal website, Power BI lists team sites within the institution that the developer can publish the data portal. The developer must be a member of these team sites and can invite staff members with an institutional email address. Obtaining a public web address to the data portal requires added administration to ensure exclusion of any personally identifiable information. Griffith University has cybersecurity policies that closely guard against the release of personally identifiable information and has a site license for Power BI, so the cost is shared and the effort of publishing the data portal is reduced. This approach supplies economies of scale and load balancing of data portal usage.

Figure 3: Power BI development environment, publication pathways, and costs



5.3 Relational Data Mining – Power BI Insights

Power BI has a collection of automated algorithms that perform relational data mining called 'Insights'. Power BI has two methods to generate insights: (1) on the whole relational database, or (2) by individual tiles[16] that include tables, and graphs embedded within a report. Importantly, these insights are outputted within a new tile that could supply further insights, allowing researchers to drill down into the database. The insights algorithms are disabled in the current public version of data portal, as is the ability to download data. In lieu of dynamically generating insights, sample insight tiles can be found on the data portal's last tab labelled 'Insights'. The types of insights supported by Power BI include:[17]

- Category outliers
- Change point in a time series
- Correlations
- Low variance
- Majority factors
- Outliers
- Overall trends in times series
- Seasonality in time series
- Steady share
- Time series outliers

5.3.1 Improving Power BI's relational data mining algorithms performance

The Insights algorithms are unconfigurable and automated. Microsoft advises,[18] "Duplicate data takes valuable time away from searching for meaningful patterns." So, to increase data mining performance, hide duplicate columns or unhide interesting columns in tables because

[16] View data insights on dashboard tiles with Power BI: https://learn.microsoft.com/en-us/power-bi/consumer/end-user-insights
[17] Types of insights supported by Power BI: https://learn.microsoft.com/en-us/power-bi/consumer/end-user-insight-types
[18] Optimise your data for Power BI Quick Insights: https://learn.microsoft.com/en-us/power-bi/create-reports/service-insights-optimize

the Insight algorithms only examine unhidden columns. This ad hoc solution works for this case study. Section 5.7 discusses permanently improving the performance of data mining and other advanced analytical techniques by addressing the case study's duplicate data, data redundancy and other relational issues hampering data mining.

5.3.2 Relational data mining performs issues beyond the hiding columns solution.

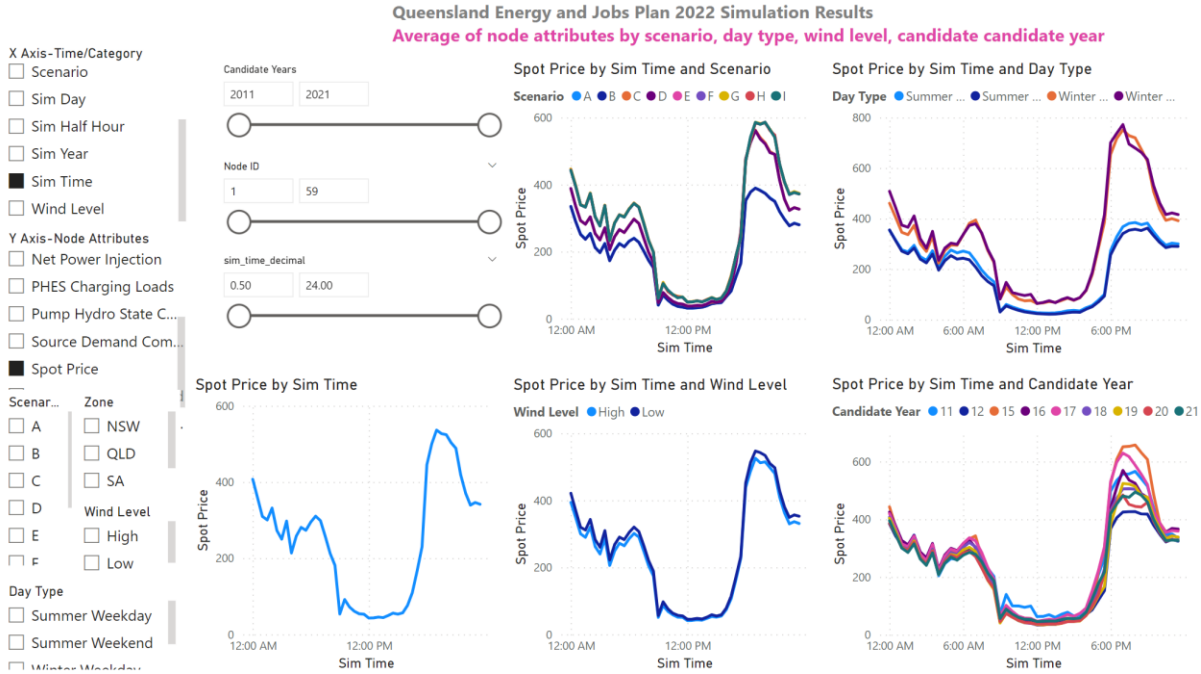
The following three data mining performance issues are unamenable to the hiding columns solution.

- Some scenarios show trivial difference between them. This means that the algorithms are searching through a billion records to find patterns when there is little to find. These large low information datasets resulted in the algorithms timing out with exhaustion. Early investigation into whether any added scenario increases information rather than just increasing data could aid in the decision to end the scenario. Early exclusion of low information scenarios would reduce the ETL time of repetitive data and support better data mining results. This screening could have also helped reduce the simulation runtime of six months.
- The 'generator half-hourly' table is over a billion records, and its 'generator details' table lacks category details about the generators. This lack of categorical detail would contribute to algorithm exhaustion. Section 6.2 further discusses this and other documentation issues.
- The lack of input data within data portal's relational database relational will also reduce the potential of relational data mining. This issue also adversely affects hybrid machine learning and simulation modelling, as discussed in Section 5.8.

5.4 Standard routine reporting versus what-if analysis using dynamic graphing

There is a place for routine reports, including standard graphs. However, researchers and policy analysts may want to make ad-hoc what-if queries and follow up on interesting relationships discovered by relational data mining. So, rather than producing a new report and graph for every situation, developing graphs with dynamic x and y-axes offers a way to easily investigate ad-hoc what-if research questions. In turn, this allows flexibility for the policy analyst to assess options easily and quickly. Example shown in Figure 4.

Figure 4: Dynamic X & Y axis options shown on two top left-hand panels. The solar duck on the five right panels shown the average NEM spot prices, by scenario, by day type, by wind level, and by candidate year.



(Source: 'Dynamic Axes' tab on the QEJP Data portal)

The ability to build a comprehensive dynamic graph relies on the underlying data being in a relational database. Familiarity with DAX would be helpful to follow a 3-minute explanation.[19] The steps involve assigning a measure to the attributes of interest. The built-in measures include the basic statistical functions, with the possibility to develop user-defined measures. Applying the same measure to all the numeric attributes is possible. Allocate the attribute measures to a Field Parameter labelled y-axis that auto-creates a slicer/filter. Allocate the categories to a Field Parameter labelled x-axis that also auto-creates a slicer/filter. Position these Field Parameters X-Axis and Y-Axis into Graph Tiles' x and y fields.

5.5 Leveraging existing normalise data to populate Geographic Information Systems

Power BI reports can also embed ESRI ArcGIS maps within tiles to display locational information, for an example, see Figure 5. ESRI's automated location of longitude and latitude for placenames does not work inside the Power BI tile when posting to a public website. The ESRI map process for Power BI tiles requires the following extra steps:

(1) Write code to download latitudes and longitude from 'Open Street Map'[20] based on placenames.

(2) Insert new latitudes and longitudes columns in the 'node details' table shown in Figure 1.

(3) Copy latitudes and longitudes data into the 'node detail' table, and

[19] Dynamic X and Y axes implementation using DAX: https://www.youtube.com/watch?v=1eurc0EY2Xg&list=PL0hL62RHC6QHgQuMu_MWnDIpeUPKsYnud&index=7

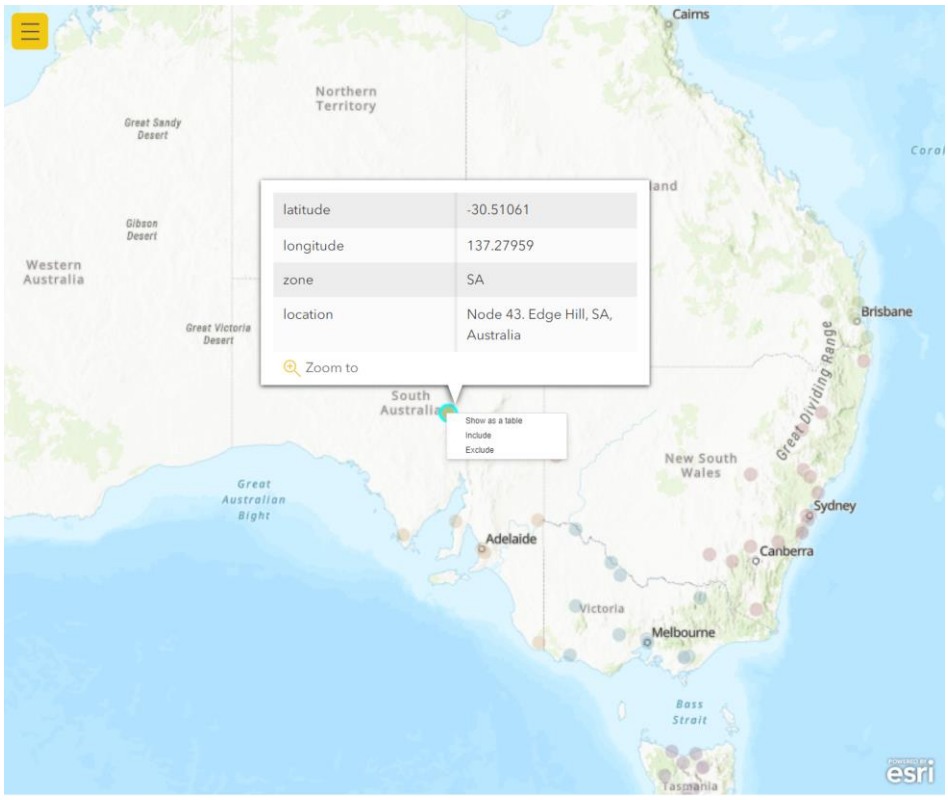
[20] Open Street Map: https://nominatim.openstreetmap.org/search

(4) Place the latitudes and longitudes fields onto a mapping tile.

Other issues encountered include: (1) Power BI's ESRI's tiles automated geolocation based on placenames failed on public websites; (2) 'Open Street Map' was unable to assign a latitude and longitude to the node named 'South East South Australia' that was renamed Mt Gambier and (3) Gaining access to use ESRI within Power BI was most challenging, having to work through three levels of security access from ESRI, Power BI, and my institution.

In addition to incorporating the latitudes and longitudes into the 'node details' table, the full functionality of the map relies on the underlying relational database. For instance, creating animations from the half-hourly datasets rely on the one-to-many relationship between the entity details and half-hourly tables. Embedded ESRI ArcGIS maps supply animation functionality.[21] This functionality can be extended to transmission flows.

Figure 5: ArcGIS map embedded in Power BI showing the nodes in the case study's simulation model.



(Source: 'Nodal Map' tab on the QEJP Data portal)

5.6 Maximising reuse of filters and categories using relational databases

One of the important aspects of the relational database design shown in Figure 1 is the consideration of where to place filters for graphs and tables. A well-designed relational database allows the placement of filters that affect the entire database. For instance, a filter in the table 'sim year' will select that year for every record in the database. This allows the

[21] Embedded ESRI ArcGIS maps supply animation functionality: https://www.youtube.com/watch?v=CywPHwK1-1c&list=PL0hL62RHC6QHgQuMu_MWnDIpeUPKsYnud&index=2

reuse of the same filter throughout the entire data portal. This reusability is timesaving and reduces the possibility of errors. Correctly, defining the filters is also essential for the functionality of dynamic graphs and embedded ESRI map.

5.7 Reducing workflow time and disk space and increasing relational data mining and other advanced analytics performance.

The simulations in this case study took 6 months to run and extracting, transforming, and loading the data into the portal and its design took 3 months. How to reduce this 9-month workflow? Developing an ERD, and normalising data are two business data modelling techniques that offer a clear view of the data structure, so instructive in cutting data redundancy. They also act like a prism to shed new light on improving the simulation model's processes. Four areas for improving the predictive data model include cutting data redundancy; geographically aligning the output data; aligning the data temporally; and breaking down the composite entities into their components. Each of these is discussed below.

5.7.1 Cutting data redundancy using calculated field

- The entity 'transmission line half-hourly' attributes 'Branch Flow' and 'Branch Marginal Losses', and the node attribute 'Marginal Losses' have AEMO variants, see Figure 1. These AEMO variants have different directional flows. These attributes are redundant, as the AEMO variants are simply negative if the AEMO line flow differs from the simulation model's default direction. A vector of positive and negatives ones in the 'Line Details' and 'Node details' tables could hold all the information needed. Making this change could cut outputting and processing 594 files (3/42×8,316) and make the AEMO flow direction information explicit. If required, the Power BI supplies calculated fields for such simple relationships.
- The four Load Serving Entity (LSE) attributes 'PHES Charging Loads', 'Pump Hydro SOC', 'Storage Charging Loads', and 'Storage SOC' all have a summed by node variant. Relational databases are well-designed to sum over the one-to-many relationship between node and LSE. Making this change could cut outputting and processing 792 files (4/42×8,316) without any loss of information.

5.7.2 Aligning output data geographically

The case study's simulation model balances electricity supply and demand every half hour, using flow constraints on transmission lines to balance the electricity supply from the generators and the demand from LSEs. These three entities lines, LSEs and generators are related via nodes, see the ERD in Figure 1. Therefore, any other entity is an aggregate calculated within the simulation model. For instance, the six NEM attributes prefixed with 'min Total Variable Cost' are state and NEM aggregates of the generator's 'min Total Variable Cost'. Rewriting the simulation model's code to output a single generator 'min Total Variable Cost' table would open research questions on minimum variable costs by generator and enable relational data mining. Making this change could also cut outputting and processing 990 files (5/42×8,316) while gaining information.

5.7.3 Aligning output data temporally

The case study's simulation model balances electricity supply and demand every half hour in this case study. Therefore, any period other than half-hourly, such as the daily 'Carbon Emissions' by generator, is an aggregate calculated within the simulation model. Rewriting the simulation model's code to output the carbon emission very half hour would allow

combining carbon emissions with the other half-hourly generator attributes and open research questions on emissions and time of day effects and improve relational data mining.

5.7.4 Breaking down composite entities into its components

The inequality multiplier entity has three attributes, 'Inequality Multiplier Upper', 'Inequality Multiplier Lower', and 'Inequality Multiplier'. These three attributes are shadow prices associated with relaxing the upper and lower constraints and their difference for the lines and generators. Hence, splitting the inequality multiplier entity between the line and generator entity could eliminate it as an unnecessary entity. The inequality multiplier entity's three attributes 'Inequality Multiplier', 'Inequality Multiplier Upper', and 'Inequality Multiplier Lower' would become attributes of the line and generator entities to enable relational data mining and simplify analysis more generally. The size of the CSV file for the inequality multiplier is 100 GB. The inequality multiplier as an entity was omitted from the data portal because it would double the size of the portal's binary file, increase the portal's latency, and shadow pricing may hold little interest for most people.

5.8 Using machine learning on simulation modelling results of high-frequency parameter sweeps to find optimal development pathways.

The literature review discusses hybrid machine learning and simulation modelling as an approach to solve computationally intensive issues such as modelling optimal development pathways at sufficiently high frequency to provide realistic modelling of storage and generation to include those relying on high-frequency settlement for financial viability. The lack of input data in the data portal discussed in Step 6 of Section 4 and the other data redundancy, and geographic and temporal misalignment discussed in Section 5.7 make the current platform's database unsuitable for this hybrid technique.

6 Other Costs

As discussed, the time to complete the ETL and develop the was 3 months and to complete the simulations was 6 months. These costs are likely to decrease with each iteration of the data portal and rewriting the simulation model to incorporate the insights gathered from the business data modelling. Entities within a principle-agent relationship can perceive the following other costs differently and adopt different strategies.

6.1 Hosting and software

The university has a site license for Power BI that allows publishing Power BI data portals to the web. There is also the cost of Premium BI development environments that allow developers to publish to group sites and the web. The university distributes these costs across the university, and a shared site allows the balancing of access load. For those working where the employing organisation does not have a site licence, the cost of Premium Power BI needs to be considered.

6.2 Documentation and Principle-Agent problem.

Documentation comes at a cost but enables the public or other researchers to use the simulation model. Publishing to the web forces a more disciplines approach to developing documentation.

There are two forms of documentation to consider, the first category of documentation is that essential to the functionality of the relational database, and the general second reference material found in PDFs or other static media. The case study has inadequacies in both forms of documentation. For instance, under the first category of documentation, functionality: the generator details table lacks attributes for generator type. Lacking this attribute hinders

relational data mining. Under the second category of documentation, general reference, the case study's simulation model is a modified version of the US electricity system.

The documentation issue is related to the principle-agent problem and ETL strategy discussed in Step 1 of Section 4. The best strategy as a self-interested individual researcher or agent is to supply only sufficient documentation to publish in a journal. This approach is consistent with strategy 1. The best strategy for an organisation or principle paying for the researcher is to have complete documentation that is consistent with Strategy 2.

7 Discussion

The energy transition presents an enormous modelling challenge to inform policy and investment decisions during the next 30 years. Researching and modelling the NEM will become more challenging during the energy transition and would benefit from big data analytics, standardisation, and specialisation to gain economies of scale. Numerous drivers are contributing to the increasing challenge, including the shift from large, centralised baseload load and peaking generation and mono-directional T&D designed to geographically dispersed VRE and renewable storage where the distinct between generators and consumers has become blurred requiring a bidirectional T&D design. To help meet this challenge, this paper argues and outlines the use of data analytic platforms to help share modelling results and enable advanced analytics on the modelling results. Five approaches to amplify the benefits of data analytics platforms to help meet the energy transition modelling challenge include, increasing economies of scale, building on open-source models, model averaging, rapid analysis, and iterative design (RAID), integrating simulation model and analytics platform workflows.

7.1 Increasing research economies of scale to meet the energy transition modelling challenge.

Meeting the challenge of modelling the energy transition requires the ability to share costs and the willingness to work collaboratively within large teams to contribute to major long-term research programs to shape future generations. Compared to other disciplines, economics has a relatively small number of researchers in a project. For example, the pure economics paper with the largest number of authors is 17 (Coenen et al., 2012). The economics paper with the largest number of collaborative research is 22 (Benjamin et al., 2012). The physics paper with the largest number of authors exceeds more than 5,000 authors. This situation indicates the potential for a more industrial-sized workflow, specialisation, standardisation to gain economies of scale. Whether finding a more precise estimate of the size of the Higgs boson or optimal development pathway, both require appropriate social and technical infrastructure.

One of the less glamorous aspects of fermenting the industrial revolution was the development of standards to allow the purchase of standard parts to develop new products. For instance, standardised bolts. Researchers modelling the energy transition bolt together datasets from different sources, including meteorological and energy. The time spent extracting, transforming, and loading these datasets to merge them could be better used analysing the results. Embedding the BoM and AEMO as part of a research value added chain or workflow could provide benefit to modelling the energy transition. Some simple value gains include transforming the datasets at the BoM and AMEO by using Excel's maximum record limit of one million records to compose yearly datasets rather than monthly. More ambitious value gains include the BoM providing weather station datasets with interpolated values in addition to the raw datasets. Another ambitious value-added gain

includes the AEMO and BoM providing their datasets within SQLite, a free open-source database management system, that has a maximum database size of 281 terabytes, a theoretical maximum number of rows per table of 2^{64} (1.8×10^{19}), default maximum number of columns per table of 2,000, and maximum number of tables per database is a little over 2 billion.[22]

7.2 Repeatability, predictability, sound argument, and model commercialisation

If the results of simulations are not replicable, any policy decisions based on the simulations may be misinformed. Sharing the input data and results with other researchers offers closer scrutiny and detection of errors, potentially prevents misinformed policy decisions. The principle-agent problem arises, as it can be in the interest of the researcher to keep input data or simulation model source code undisclosed.

Similarly, some commercial simulation model approaches are at odds with academic expectations of repeatability, predictability, transparency, and sound argument. The interests of the commercial supplier of simulation models are served by non-disclosure of source code and input data. The NEM models have many variables, enabling tuning the model to fit the data. Analogous to first order predicate logic, a true conclusion requires both a true premise and sound argument, a simulation model's true output requires both accurate input data and sound assumptions in the code. A well-tuned model may well provide high predictive performance for small changes in variables, where the 'arguments' in the code are conditionally valid for small changes in variables. However, the energy transition is anything but a small change in variables. This situation makes the ability to check the source code for sound argument important. Noting, open-source code that would allow checking for sound argument is not synonymous with free. For example, neither Energy Exemplar nor IES's NEM model are free for commercial users. IES's open-source code approach contrasts with Energy Exemplar's closed source-code approach. Iowa University's AMES model is both free and open source, except for the inbuilt QP optimiser from IBM. However, AMES, unlike Prophet and PLEXOS, only models DC transmission flows and lacks a standard NEM baseline configuration.

7.3 Selecting the 'best model' versus averaging models and retaining information

The standard econometric approach is to select the 'best model' that comprises a 'best fit' weighted by some penalty for increasing number of variables. This approach works well under certain conditions but is ill-suited to the energy transition that has no historical precedent to find a 'best fit' and with the NEM models having many variables that allow finetuning the models for the current NEM structure. A structure that will cease to exist during the transition.

Model averaging provides an alternative approach to model selection. Bates and Granger (1969) introduce 'model-averaging' to improve forecasting accuracy. Model averaging involves using the same input data in different simulation models and averaging the different models' outputs. Clemen (1989) reviews the combining forecasts literature and concludes that (1) combining multiple individual forecasts improved forecast accuracy, and (2) simple combinations of models often work well, compared to more complex methods. His review discusses combining differing models to improve forecast accuracy, or 'model-averaging'. Model-averaging has an extensive literature (Fernández, Ley, & Steel, 2001; Garratt, Lee, Mise, & Shields, 2008; Garratt, Lee, Pesaran, & Shin, 2003; O'Hagan, 1995). In addition to

[22] SQLite database limits: https://www.sqlite.org/limits.html

forecast accuracy, other advantages over model selection include retaining information and more robust confidence intervals.

Ideally, equal-weighted-model-averaging AMES, Prophet and PLEXOS would provide more robust scenario analysis. However, the preference for open-source models to check arguments would discount PLEXOS's inclusion. Furthermore, PLEXOS neither replied to a query about academic licensing nor provided any academic licensing information on its website. Consequently, PLEXOS was excluded from planned model averaging and comparative analysis. AMES is free for academic and commercial use; however, its optimiser is only free for academic use. Prophet is free for academic use; however, its high-performance LP optimisers require commercial licenses. In summary, model averaging AMES and Prophet could offer more robust confidence intervals and testing different model averaging techniques using Power BI's calculated fields would be straightforward, as would be other comparative analysis.

7.4 Applying Rapid Analysis and Iterative Design principles to scenario selection.

The case study's simulations took 6 months to complete, and the development of the data analytics platform required a further 3 months. The data mining algorithms looking for relationships in the results' data timed-out with exhaustion.

Cutting data redundancy is one method to reduce workflow time and improve data mining performance, discussed in Section 5.3. This issue is amenable to re-coding the simulation model to output tables in a different structure or omit and use calculated fields in the analytics platform.

Cutting repetitive or similar scenarios is another method to reduce workflow times and improve data mining performance, discussed in Section 5.3.2. This issue is not amenable to re-coding the simulation model, but is amenable to applying Rapid Analysis and Iterative Design (RAID) Principles (Wilson, 1999) using an analytics platform.

RAID is the predecessor to Agile. RAID uses early prototyping to gain feedback to change design iteratively. Developing the analytics platform before the simulations supplies the ability to perform sensitivity analysis on each new simulation and if a scenario lacks any significantly different impact to other scenarios, exclude it from the project. This approach could help reduce workflow time and data mining exhaustion and allow design changes to investigate other more interesting scenarios. RAID principles to scenario selection could be further extended by using Prophet's faster LP to select scenarios rather than AMES's slower QP. Furthermore, the application of the template effect discussed in Section 5.1 finds a good fit with the RAID principles.

7.5 Iterating the data portal and integrating the AMES simulation model

Prophet and PLEXOS have sophisticated scenario parameter sweeping, comparative scenario analysis, and integrated analytics platforms. These features are lacking in the original AMES that produces a single output file per simulation input file intended for manual analysis. The modified NEM version of AMES used in this case study produced 8,316 output files. To bring AMES closer to the functionality of Prophet and PLEXOS, three major factors need addressing, including (1) automated generation of parameter sweep input files, (2) automated cleaning, transforming, loading of simulation output files, and (3) integration into a data analytics platform.

7.5.1 Automating the generation of parameter sweep input files.

The case study's simulations were concluded before the data portal was envisioned. As discussed in the Section 4 (Step 6), there is massive data redundancy in the input files, making their inclusion in the data portal too expensive. Automating the generation of AMES input files would reduce data redundancy and enable their inclusion of the input data in the portal to enable more extensive machine learning and other analytical techniques. Some pre-project planning is required to realise the full potential of the data portal's advanced analytics.

Given the data analytic platform, Power BI, requires the input data in a relation database format, generating the AMES's simulation input data files from a relational database could reduce workflow time and handling errors. Furthermore, SQLite claims that it can read and write data about 35% faster than the underlying file system.[23] The auto generation of parameter sweeps was discussed in Section 4 (Step 6). Other workflow options require consideration.

7.5.2 Automating the cleaning, transforming, loading of simulation output files.

Given the destination of the AMES's output files is a relational database in Power BI, recoding the AMES model to output its simulation tables directly to an SQLite would eliminate post-processing cleaning and a transformation step. For instance, for the case study, this recoding would reduce the need to clean and transform 8,316 tabulated text output files distributed among the terminal branches of a three-levelled branching hierarchy directory structure into 198 database files within a single directory. This recoding would reduce logistics with the added advantage of reducing the possibility of mistakes.

7.5.3 Integrating AMES and Prophet results for comparison into a data analytics platform

The case study has illustrated the low technical threshold to complete a data analytics platform using Power BI and discussed the ample scope for further development. The benefits of the relational database underpinning Power BI include easier model comparison, model averaging, and hybrid simulation modelling and machine learning. This approach is standard in meteorology where in cyclone tracking, models from several countries are used to model cyclone pathways and then compared to provide the average or most likely pathway.[24] Completing the data analytics platform was useful in determining the potential for Power BI to aid researchers in the energy transition and identifying strengths and weakness in the AMES model.

8 Conclusion

There is a significant upfront cost in time in developing a relational database and data portal for a simulation project's results; however, the advantages of relational databases and portals are many. Pre-project planning and agreement on ETL strategy is needed to realise the full potential of the advanced analytics of using a data portal and avoid principle-agent problems. Four forms of benefit include: (1) The immediate benefit from the reuse of the data to answer further research questions, whether by the original researcher or by other researchers. This reuse may increase as researchers become more accustomed to using other's simulations results. (2) The development of the relational database and data portal

[23] SQLite 35% faster than the underlying file system: https://www.sqlite.org/fasterthanfs.html
[24] World Meteorology Organisation: Tropical Cyclone Forecaster Websitehttps://severeweather.wmo.int/TCFW/

becomes a feedback mechanism to supply insights into improving the simulation model. (3) The upfront cost per project will decline given learning curve and template effects and separating the analytical and simulation modelling platforms. (4) Transforming simulation output results into a relational database allows reanalysis with other techniques such as relational data mining and machine learning.

9 Acknowledgements

Thank you to Dr P. Wild for supplying the case study's simulation model output results. Thank you to the leader and participants in a focus group who supplied feedback on improving the data portal, Isaac Jennings, Lucia-Valentina Stan, Shell Energy, Clean Co Queensland, Powerlink Queensland, Department of Energy and Public Works, Queensland land Government.

10 References

AEMO. (2021). *Integrated Systems Plan Methodology*. Retrieved from https://aemo.com.au/-/media/files/major-publications/isp/2021/2021-isp-methodology.pdf?la=en

ARENA. (2023). Australian Renewable Energy Agency. Retrieved from https://arena.gov.au/

Bates, J. M., & Granger, C. W. J. (1969). The Combination of Forecasts. *Journal of the Operational Research Society, 20*(4), 451-468. doi:10.1057/jors.1969.103

Bell, W. P., Wild, P., Foster, J., & Hewson, M. (2017). Revitalising the wind power induced merit order effect to reduce wholesale and retail electricity prices in Australia. *Energy Economics, 67*, 224-241.

Benjamin, D. J., Cesarini, D., Chabris, C. F., Glaeser, E. L., Laibson, D. I., Guðnason, V., . . . Lichtenstein, P. (2012). The Promises and Pitfalls of Genoeconomics. *Annual Review of Economics, 4*(1), 627-662. doi:10.1146/annurev-economics-080511-110939

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting, 5*(4), 559-583. doi:https://doi.org/10.1016/0169-2070(89)90012-5

Coenen, G., Erceg, C. J., Freedman, C., Furceri, D., Kumhof, M., Lalonde, R., . . . in't Veld, J. (2012). Effects of Fiscal Stimulus in Structural Models. *American Economic Journal: Macroeconomics, 4*(1), 22-68. doi:10.1257/mac.4.1.22

de Ville, B. (2001). *Microsoft Data Mining: Integrated Business Intelligence for e-Commerce and Knowledge Management*: Digital Press.

Džeroski, S., & Lavrač, N. (2001). *Relational Data Mining*. Berlin, Heidelberg, New York: Springer-Verlag.

Earley, S., Henderson, D., & Data Management Association. (2017). *Dama-Dmbok : Data Management Body of Knowledge* (Second ed.). Bradley Beach, New Jersey: Technics Publications.

Fernández, C., Ley, E., & Steel, M. F. J. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics, 100*(2), 381-427. doi:https://doi.org/10.1016/S0304-4076(00)00076-2

Foster, J., Bell, W. P., Wild, P., Sharma, D., Sandu, S., Wagner, L., . . . Bagia, R. (2013). *Analysis of institutional adaptability to redress electricity infrastructure vulnerability due to climate change*. Gold Coast: National Climate Change Adaptation Research Facility.

Garratt, A., Lee, K., Mise, E., & Shields, K. (2008). Real-Time Representations of the Output Gap. *The Review of Economics and Statistics, 90*(4), 792-804. Retrieved from http://www.jstor.org/stable/40043115

Garratt, A., Lee, K., Pesaran, M. H., & Shin, Y. (2003). Forecast Uncertainties in Macroeconomic Modeling: An Application to the U.K. Economy. *Journal of the*

American Statistical Association, 98(464), 829-838. Retrieved from http://www.jstor.org/stable/30045334

Grozev, G., & Batten, D. (2006). NEMSIM: Finding Ways to Reduce Greenhouse Gas Emissions Using Multi-Agent Electricity Modelling. In P. Perez & D. Batten (Eds.), *Complex Science for a Complex World: Exploring Human Ecosystems with Agents* (pp. 227-254). Canberra: ANU Press.

Guevara, J., Loaiza, O. L., Lévano, D., & Zambrano, W. E. I. (2022). *Management System for Records in the Context of Business Analysis with BABOK for Native Organizations in Latin America*, Cham.

Hansen, P., Liu, X., & Morrison, G. M. (2019). Agent-based modelling and socio-technical energy transitions: A systematic literature review. *Energy Research & Social Science, 49*, 41-52. doi:https://doi.org/10.1016/j.erss.2018.10.021

Hoekstra, A., Steinbuch, M., & Verbong, G. (2017). Creating Agent-Based Energy Transition Management Models That Can Uncover Profitable Pathways to Climate Change Mitigation. *Complexity, 2017*, 1967645. doi:10.1155/2017/1967645

IIBA. (2015). *A Guide to the Business Analysis Body of Knowledge (BABOK Guide) Version 2.0* (Third ed.). Ontario: International Institute of Business Analysis.

Macariola, R. N., & Silva, D. L. (2019). Coping with the Information Age: Development of a Data Flow Diagram-Based Knowledge Management System for Mitigating Delays for Construction. *IOP Conference Series: Materials Science and Engineering, 652*(1), 012070. doi:10.1088/1757-899X/652/1/012070

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparison. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 99-138. Retrieved from http://www.jstor.org/stable/2346088

Padhy, N., & Panigrahi, R. (2012). Multi Relational Data Mining Approaches: A Data Mining Technique. *International Journal of Computer Applications, 57*(17), 0975-8887. Retrieved from https://arxiv.org/ftp/arxiv/papers/1211/1211.3871.pdf

Schimeczek, C., Deissenroth-Uhrig, M., Frey, U., Fuchs, B., Ghazi, A. A. E., Wetzel, M., & Nienhaus, K. (2023). FAME-Core: An open Framework for distributed Agent-based Modelling of Energy systems. *The Journal of Open Source Software, 8*(84). doi:https://doi.org/10.21105/joss.05087

Shinde, P., & Amelin, M. (2019, 21-24 May 2019). *Agent-Based Models in Electricity Markets: A Literature Review.* Paper presented at the 2019 IEEE Innovative Smart Grid Technologies - Asia (ISGT Asia).

Sklyar, V. (2021). *Business Analysis Learning Perspectives in ICT Education.* Paper presented at the ICT in Education, Research and Industrial Applications. Integration, Harmonization and Knowledge Transfer, Kherson, Ukraine.

Sun, J., & Tesfatsion, L. (2007). Dynamic Testing of Wholesale Power Market Designs: An Open-Source Agent-Based Framework. *Computational Economics, 30*(3), 291-327.

Takai, T., Shintani, K., Andoh, H., & Washizaki, H. (2020, 1-15 Sept. 2020). *Continuous modeling supports from business analysis to systems engineering in IoT development.* Paper presented at the 2020 9th International Congress on Advanced Applied Informatics (IIAI-AAI).

Valêncio, C. R., Oyama, F. T., Neto, P. S., & Colombini, A. C. (2012). MR-Radix: a multi-relational data mining algorithm. *Human-centric Computing and Information Sciences, 2*(4). doi:https://doi.org/10.1186/2192-1962-2-4

von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., & Garcke, J. (2020). *Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions*, Cham.

Wild, P., & Bell, W. P. (2011). Assessing the economic impact of investment in distributed generation using the ANEM model. In J. Foster (Ed.), *Intelligent Grid Research Cluster – Project 2: marketing and economic modelling of the impacts of distributed generation*. Brisbane, Australia: CSIRO Intelligent Grid Research Cluster.

Wild, P., Bell, W. P., & Foster, J. (2012). The impact of carbon prices on Australia's National Electricity Market. In J. Quiggin, D. Adamson, & D. Quiggin (Eds.), *Carbon Pricing: Early Experience and Future Prospects* (pp. 101–122). Cheltenham, UK; Northampton, USA: Edward Elgar.

Wild, P., Bell, W. P., & Foster, J. (2015). Impact of Carbon Prices on Wholesale Electricity Prices and Carbon Pass-Through Rates in the Australian National Electricity Market. *The Energy Journal, 36*(3). doi:10.5547/01956574.36.3.pwil

Wild, P., Bell, W. P., Foster, J., & Hewson, M. (2015). *Australian National Electricity Market Model version 1.10*. In W. P. Bell (Ed.), (pp. 62). Retrieved from https://espace.library.uq.edu.au/view/UQ:360894

Wilson, S. F. (1999). *Analyzing Requirements and Defining Solution Architectures*: Microsoft Press.

Zhou, Z., Chan, W. K., & Chow, J. H. (2007). Agent-based simulation of electricity markets: a survey of tools. *Artificial Intelligence Review, 28*(4), 305-342. doi:10.1007/s10462-009-9105-x

11 Advantages of Prophet over the case study's simulation model

The free academic simulation model of the NEM called Prophet is much faster and has more functionality than the case study's simulation model, including.

- switching between 5-minute, half-hourly and hourly trading periods, using tick-boxes.
- modelling expansion optimisation
- switching between supply side modelling options, such as, bid stacks, SRMC, …
- demand side: DSM and proxy for price sensitive demand
- modelling multiple trading techniques and options
- graphical user interface (GUI) to easily add and modify network structures.
- Selectable output of reports to reduce volume of outputted data
- Selectable levels of reporting aggregation.
- Selectable aggregated reporting between simulations
- macros to perform parameter sweeps without the need to create individual input files for each simulation.
- Extraction and transformation of outputted data is much easier.
- Prophet is an industry standard package that enables easier results swapping and knowledge transfer.
- extensive documentation