Modelling data with mathematics

James Foster
CSIRO Energy (Newcastle)

AAMT Webinar
23rd November 2020

Overview

- The modelling process
- Modelling using mathematical optimisation
- Using maths to gain insight into data

The modelling process

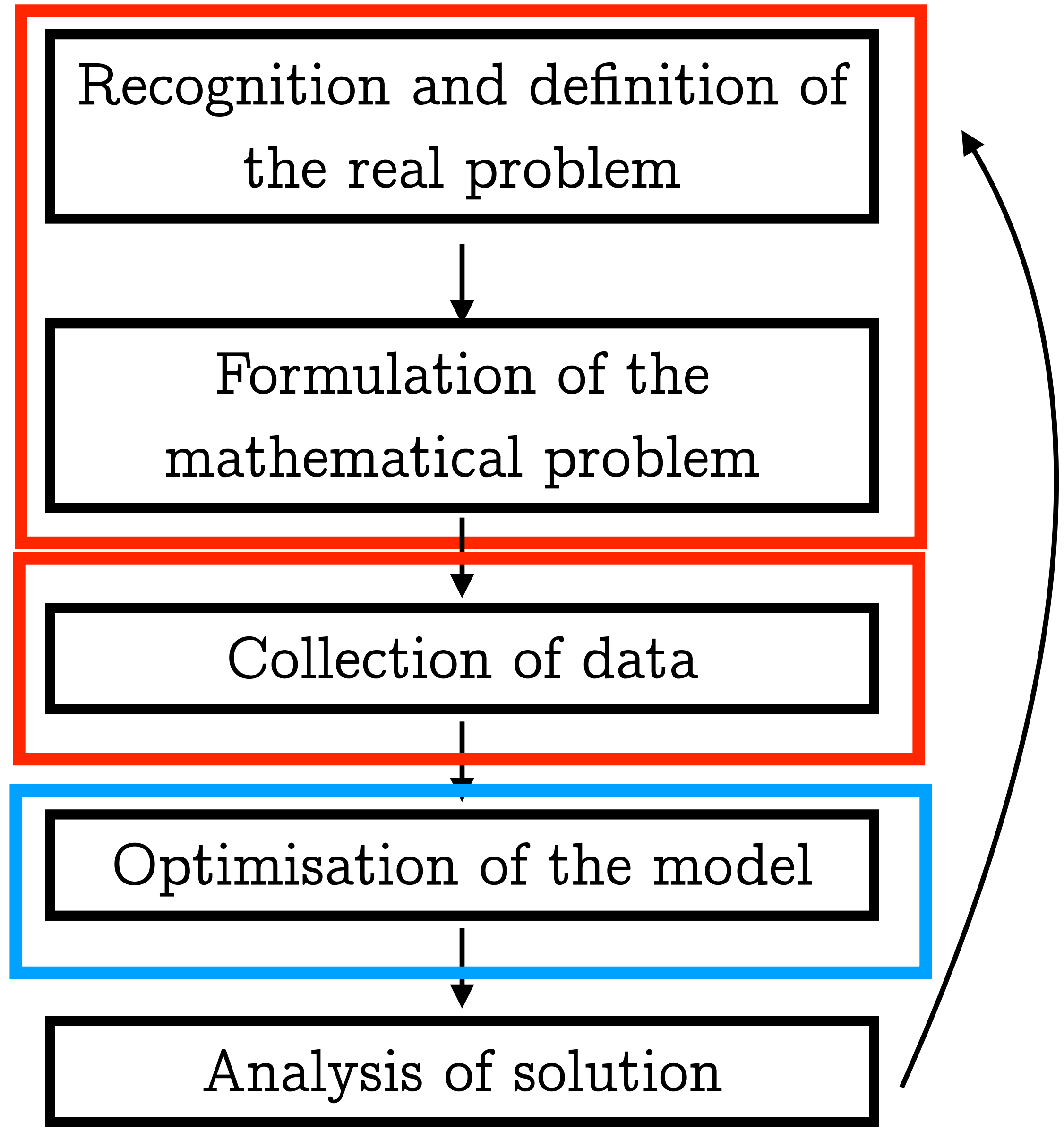
The modelling process

Task

Mathematical Modelling

Data Modelling

Solve: Algorithms + Computation



It is necessary to separate

- **model**
- **data**
- **solving**

Fragni, E., & Gondzio, J. (1999). Optimization modeling languages.

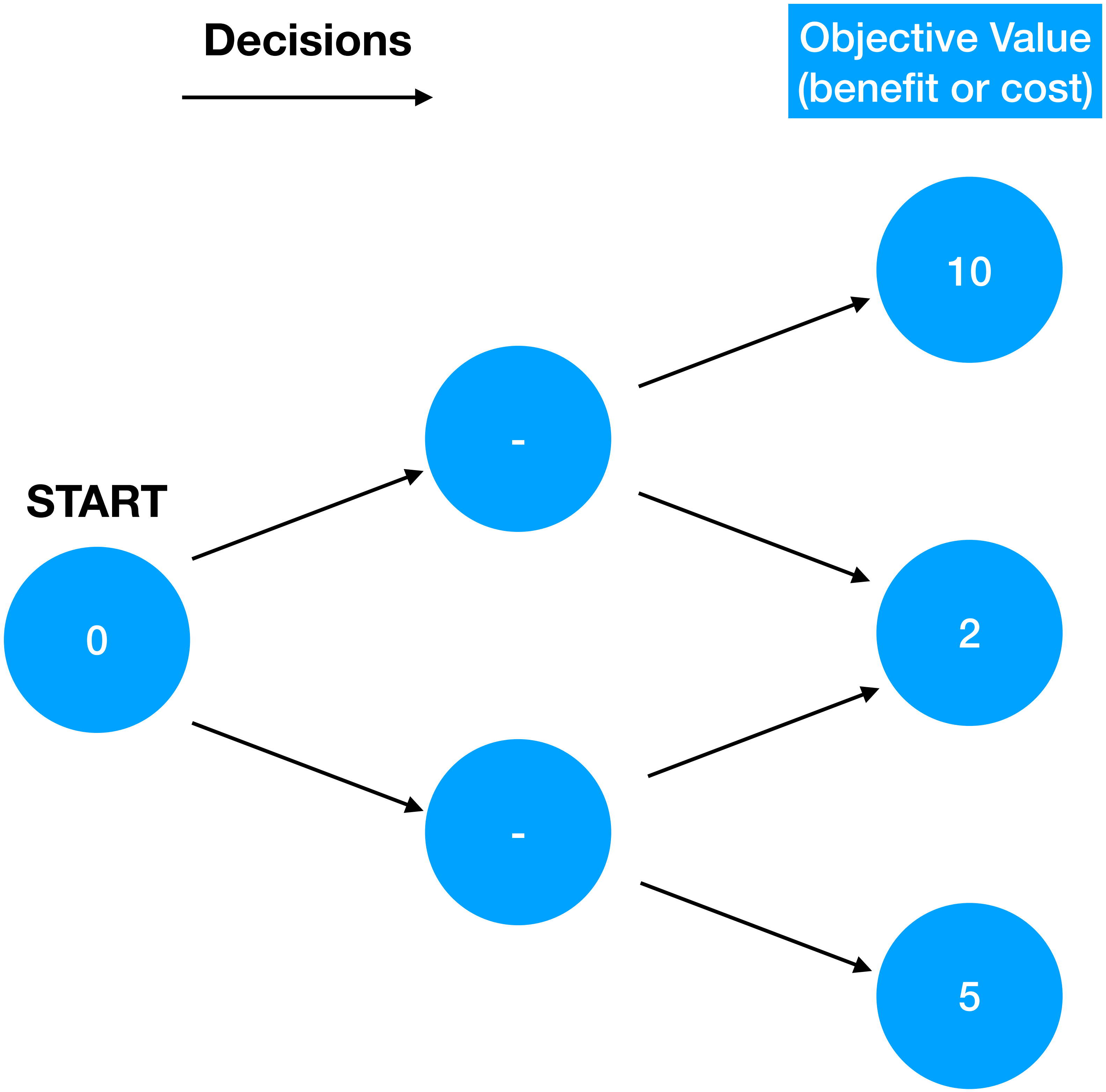
Modelling using mathematical optimisation

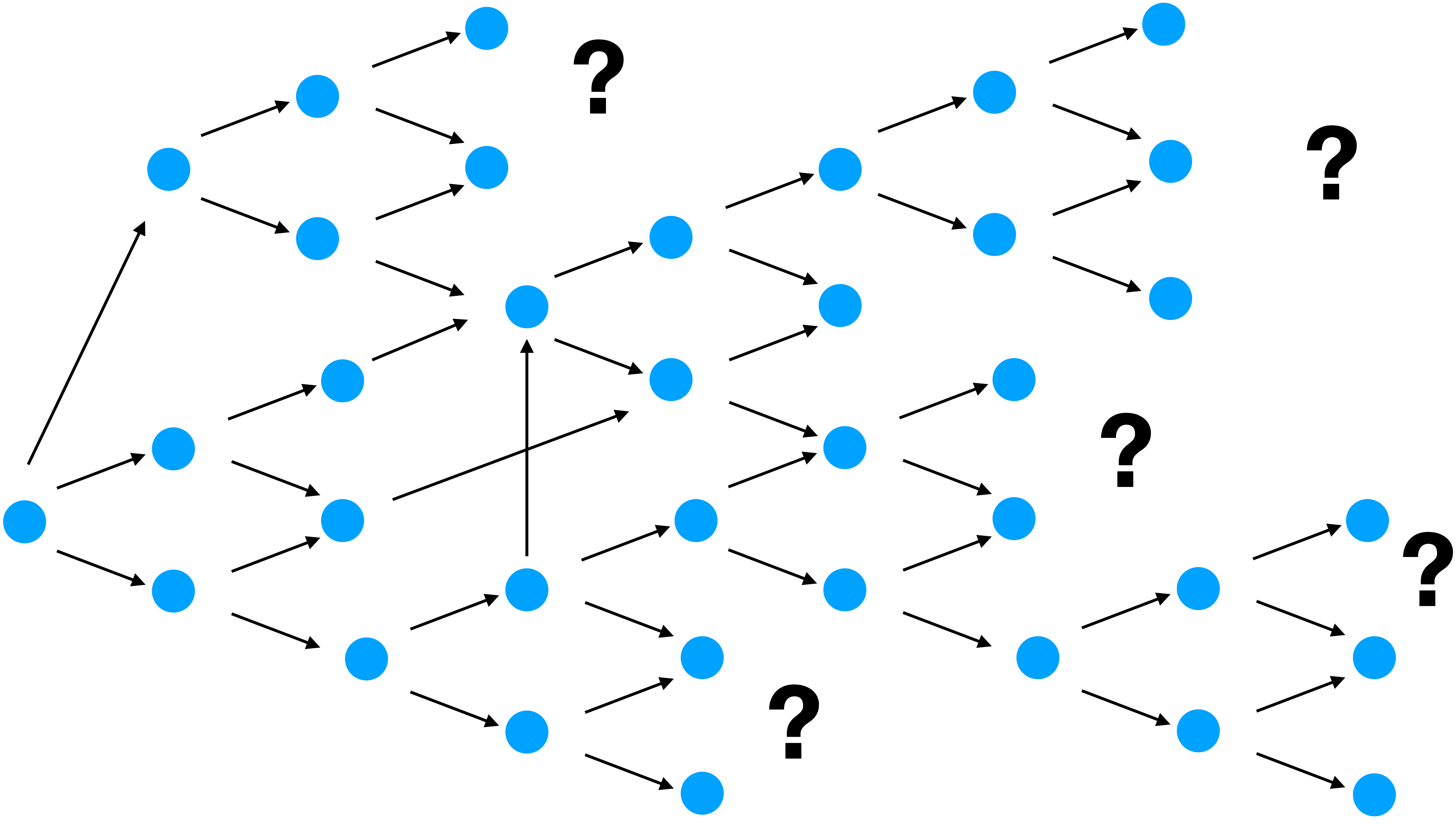
Optimisation: make optimal decisions

- **Examples:**
  - Which way?
    - Find the shortest path between two points in a network.
  - How many?
    - Allocate the best people to the right job.
  - Where do we focus?
    - Communicate with the most appropriate demographic.

Optimisation problems in power systems

- Which way?
  - unit commitment - *minimise* cost of generating power
  - optimal power flow - while considering real-world operations and physics
- How many?
  - infrastructure investment planning - *maximise* welfare
- Where do we focus?
  - state estimation - find *most likely* state of a system (e.g. a grid) to explain a set of measurements subject to measurement error


Decisions
Objective Value
(benefit or cost)
START
0
-
-
10
2
5



Why *mathematical* optimisation?

- *Without structure*, the best approach to solve an optimisation problem is to enumerate all candidate solutions
  - evaluate, score, compare, pick best
- *With* - mathematical - *structure*, things can 'simplify' a lot and apply specialised algorithms
- When we are *precise*, we can *reason*

Optimisation problems (in the abstract)

- minimise cost (function)
- subject to
  - mathematical models for entities
    - laws of physics
    - safe operation
    - financial budgets
    - resource constraints
    - technological envelopes

objective

constraint set:
Describe the situation using equations and inequalities

using linear and nonlinear functions

Mathematical structure

- linear $Ax = b$
- polynomial $x^2 + y^2 + z^2 - 2xyx - 1 = 0$
- set membership $x \in \mathbb{S}^n_+$
- integer $x \in \{0, 1\}$
- logical conditions $(x \geq 2 \wedge y = 1) \vee (x \leq y \leq 2)$

Mathematical optimisation

$$\begin{aligned} \min \quad & f(x) \\ \text{s. t.} \quad & g_i(x) \leq 0 \\ & h_j(x) = 0 \end{aligned}$$

- decision variables (vector): $x \in \mathbb{R}^n$
- functions (vector operations): $f, g_i, h_j : \mathbb{R}^n \to \mathbb{R}$
  - Their nature and type is fundamentally important
- The number of constraints and variables is typically a finite set: $i \in \{1, \ldots, m\}, j \in \{1, \ldots, l\}$

Example: the National Electricity Market

- "Big question":
  - *What is the best set of representative lines to make a network "backbone"?*
- *Sets*
  - The set of connected locations in the network (*electricity transmission substations*).
  - The set of all possible lines between the connected locations.

Example: the National Electricity Market

- *Objective:*
  - *Minimise the total length of the network.*
- *Constraints*
  - The network must be joined up i.e. not disconnected.
  - We must always join a special set of “reference nodes”

Mathematical model

- The mathematical structure is understood through
  - sets and indices (tuples)
  - parameters (given data)
  - variables (solve for these)
  - constraints (equations and inequalities)
  - objective functions (the goal)

$$\min_{x} \sum_{(i,j)\in E} c_{i,j} x_{i,j}$$

$$\text{s.t.} \quad \sum_{(i,j)\in E} x_{i,j} = \sum_{(j,k)\in E} x_{j,k} \quad \forall j \in V \setminus \{1, n\}$$

$$\sum_{(i,n)\in E} x_{i,n} = 1$$

$$0 \leq x_{i,j} \leq C_{i,j} \quad \forall (i,j) \in E$$

Dunning, I., Huchette, J., & Lubin, M. (2015). JuMP: a modeling language for mathematical optimization, 1–21. Retrieved from http://arxiv.org/abs/1508.01982

Mathematical model of a flow problem

- Minimum cost flow problem
  - over a network graph G = (V, E) = (vertices, edges)
  - V = {1,2, …n} (vertices)
  - Flow *source* at vertex 1
  - Flow *sink* at vertex n

$$\min_{x} \sum_{(i,j)\in E} c_{i,j} x_{i,j}$$

$$\text{s.t.} \quad \sum_{(i,j)\in E} x_{i,j} = \sum_{(j,k)\in E} x_{j,k} \quad \forall j \in V \setminus \{1, n\}$$

$$\sum_{(i,n)\in E} x_{i,n} = 1$$

$$0 \leq x_{i,j} \leq C_{i,j} \quad \forall (i,j) \in E$$

Dunning, I., Huchette, J., & Lubin, M. (2015). JuMP: a modeling language for mathematical optimization, 1–21. Retrieved from http://arxiv.org/abs/1508.01982

Using maths to gain insight into data

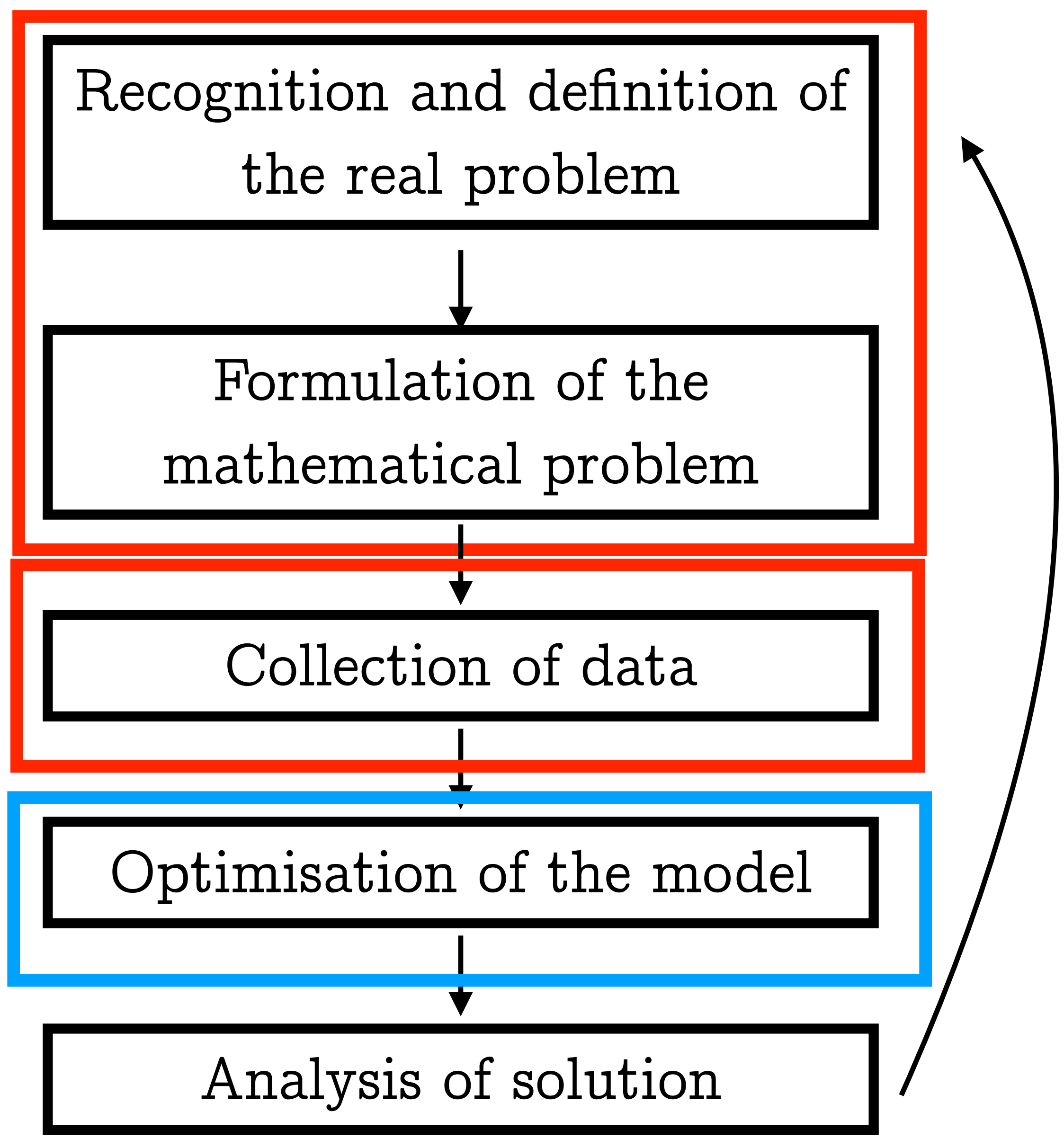
The modelling process

Task

Mathematical Modelling

Data Modelling

Solve: Algorithms + Computation

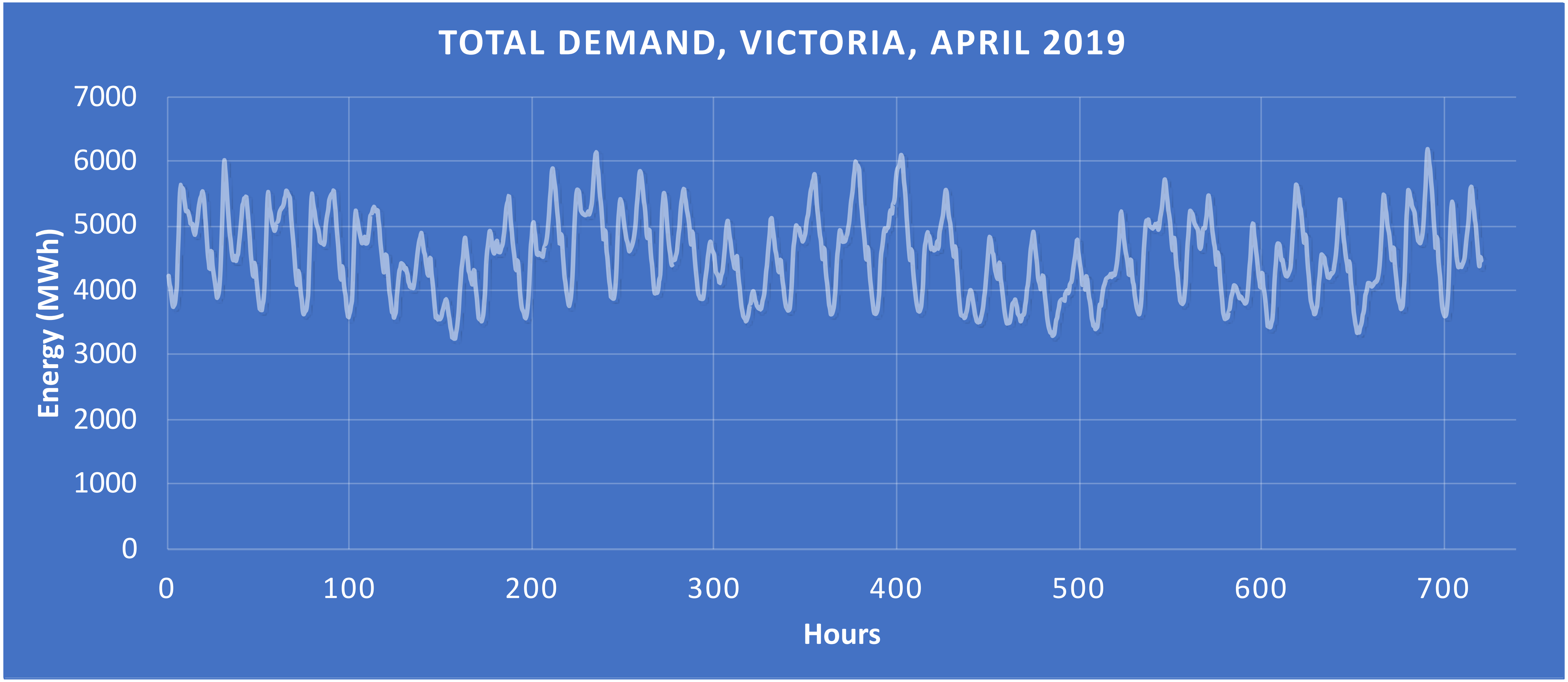


It is necessary to separate

- **model**
- **data**
- **solving**

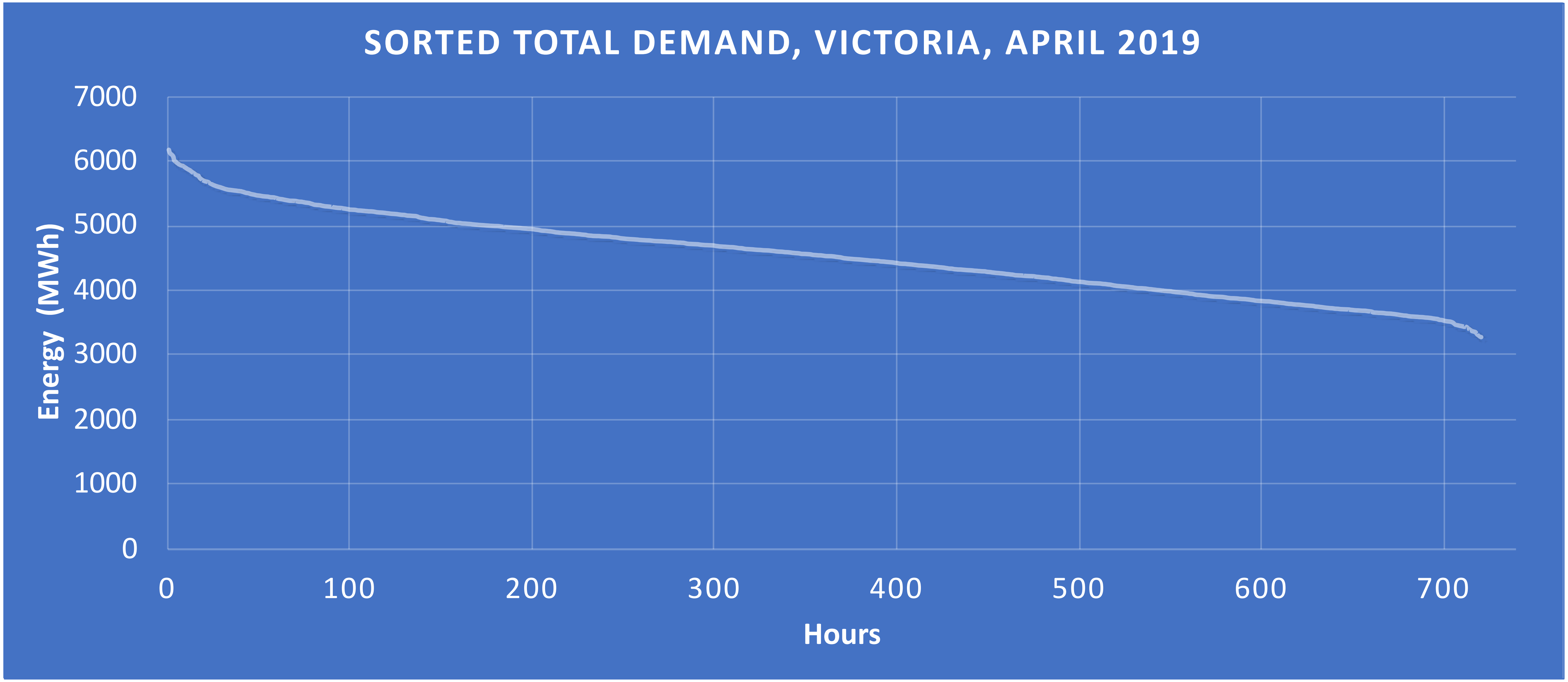
Fragni, E., & Gondzio, J. (1999). Optimization modeling languages.

Electricity demand

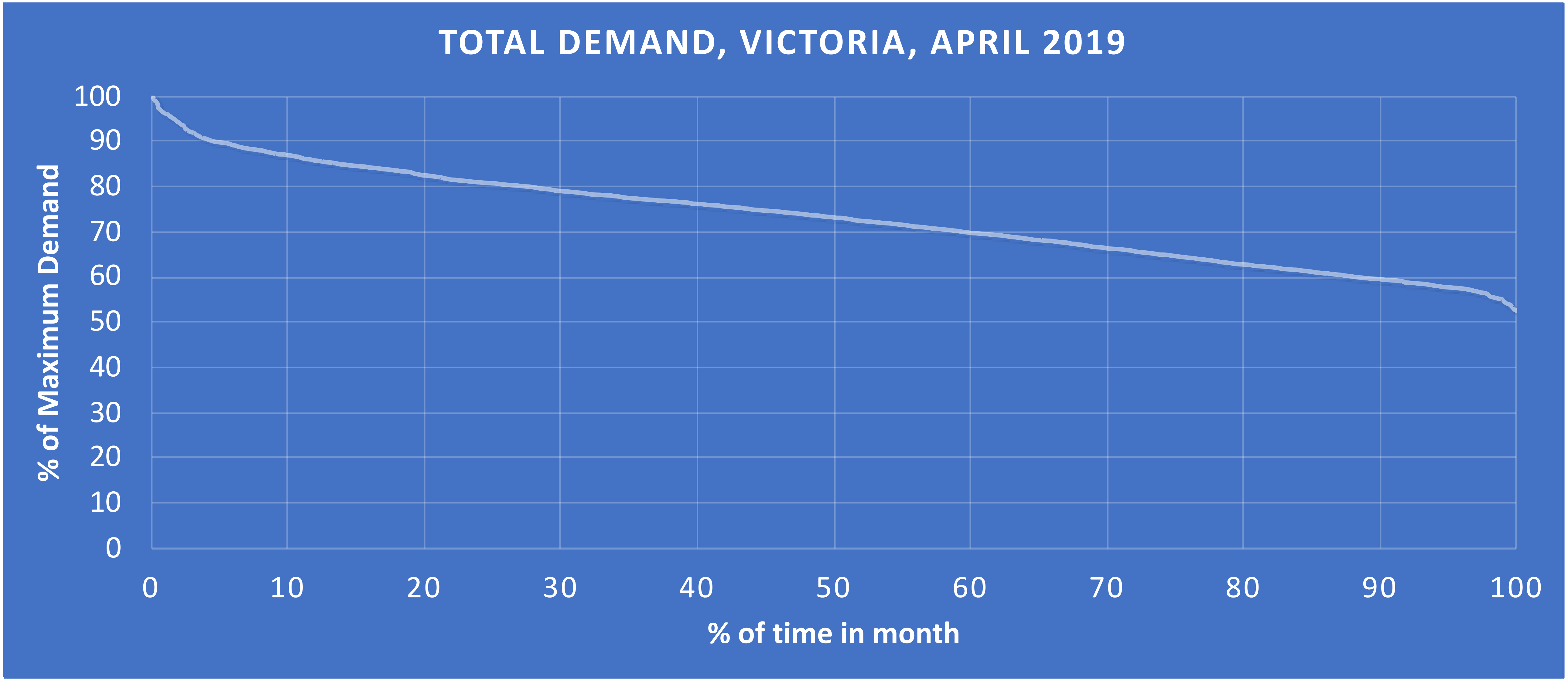


Source:
https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem

Load Duration Curves (LDCs)

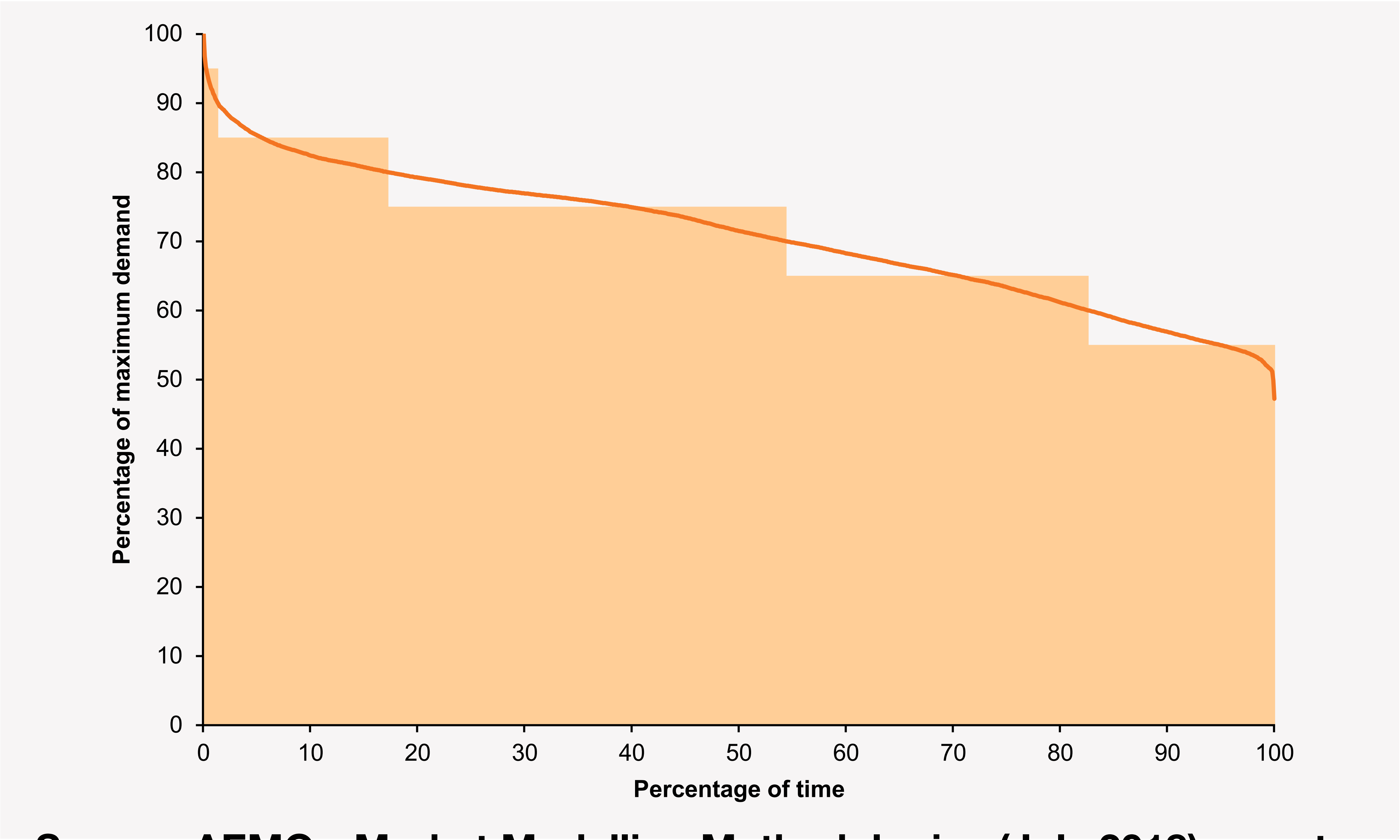


Load Duration Curves (LDCs)



AEMO Market Modelling Methodologies

Figure 8 A load duration curve partitioned into five load blocks



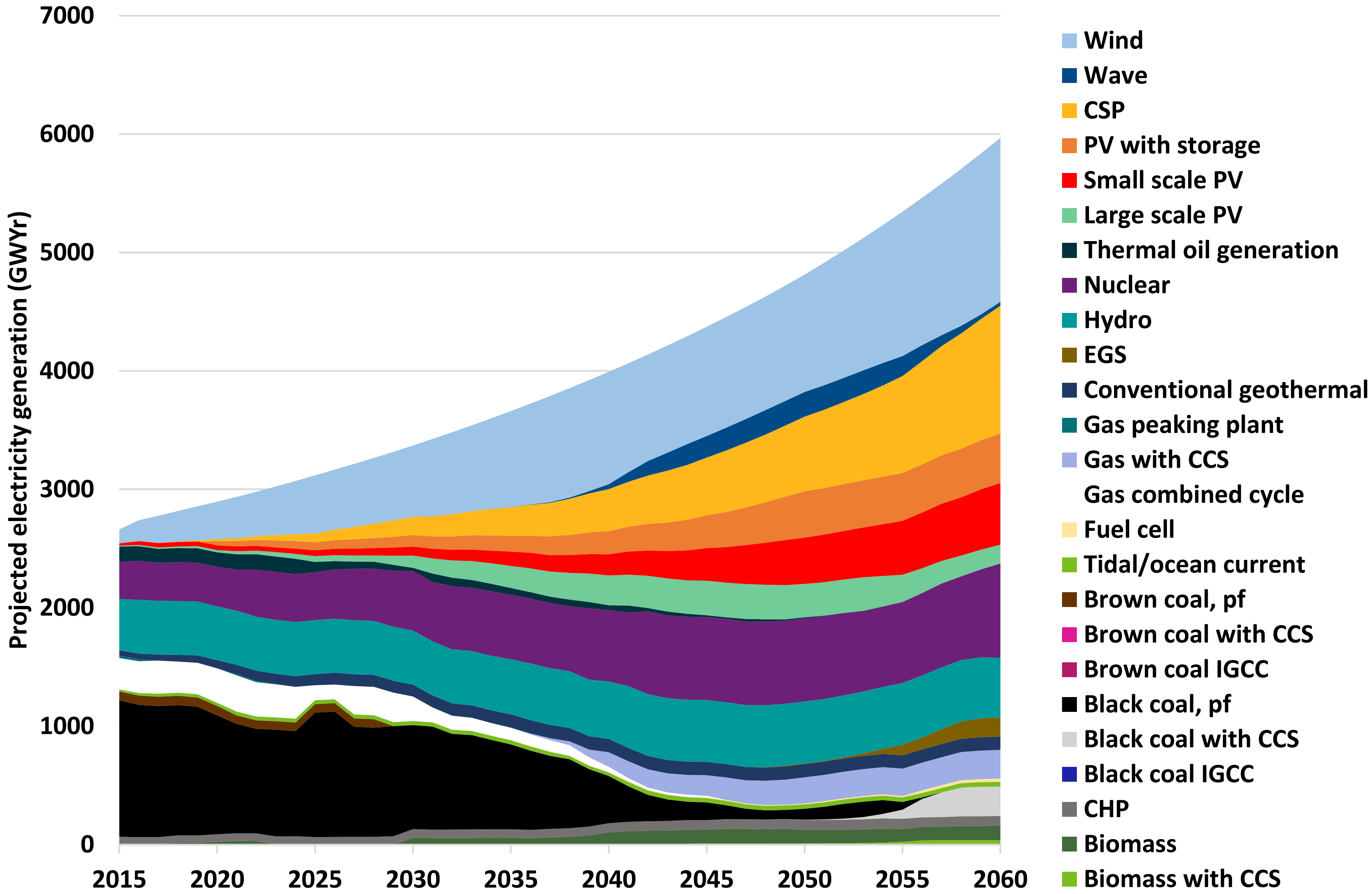
Source: AEMO - Market Modelling Methodologies (July 2018) report

AEMO Market Modelling Methodologies, page 17…

“The regional demand time series fed into the DLT is fitted with a step function so that the total number of simulation periods per day is reduced from twenty-four hours to an appropriate number of load blocks. These load blocks are created using a weighted least-square fit method which performs an **optimisation** that minimises the sum of squared errors (i.e. the square of the difference between the hourly demand fed into the model and the step function approximation). The weighted least square approach has the advantage of fitting the step function more tightly to the original demand time series – allocating more blocks to higher load periods and less to periods of low demand. The duration of each block can therefore vary depending on how the underlying intervals are grouped together. “

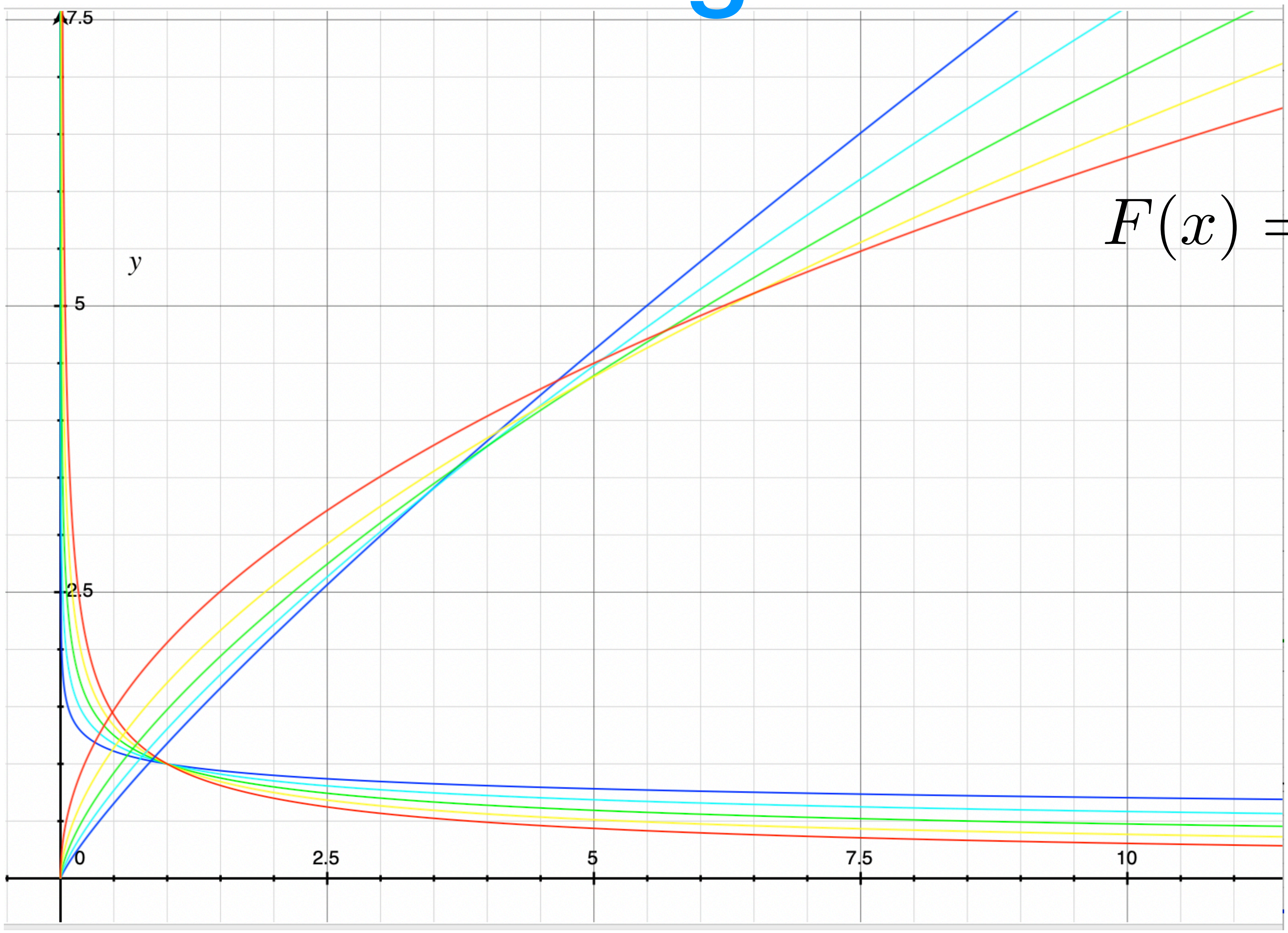
Global electricity generation by technology

2 degrees scenario (GenCost report)



Learning curves

Cost



Cumulative experience

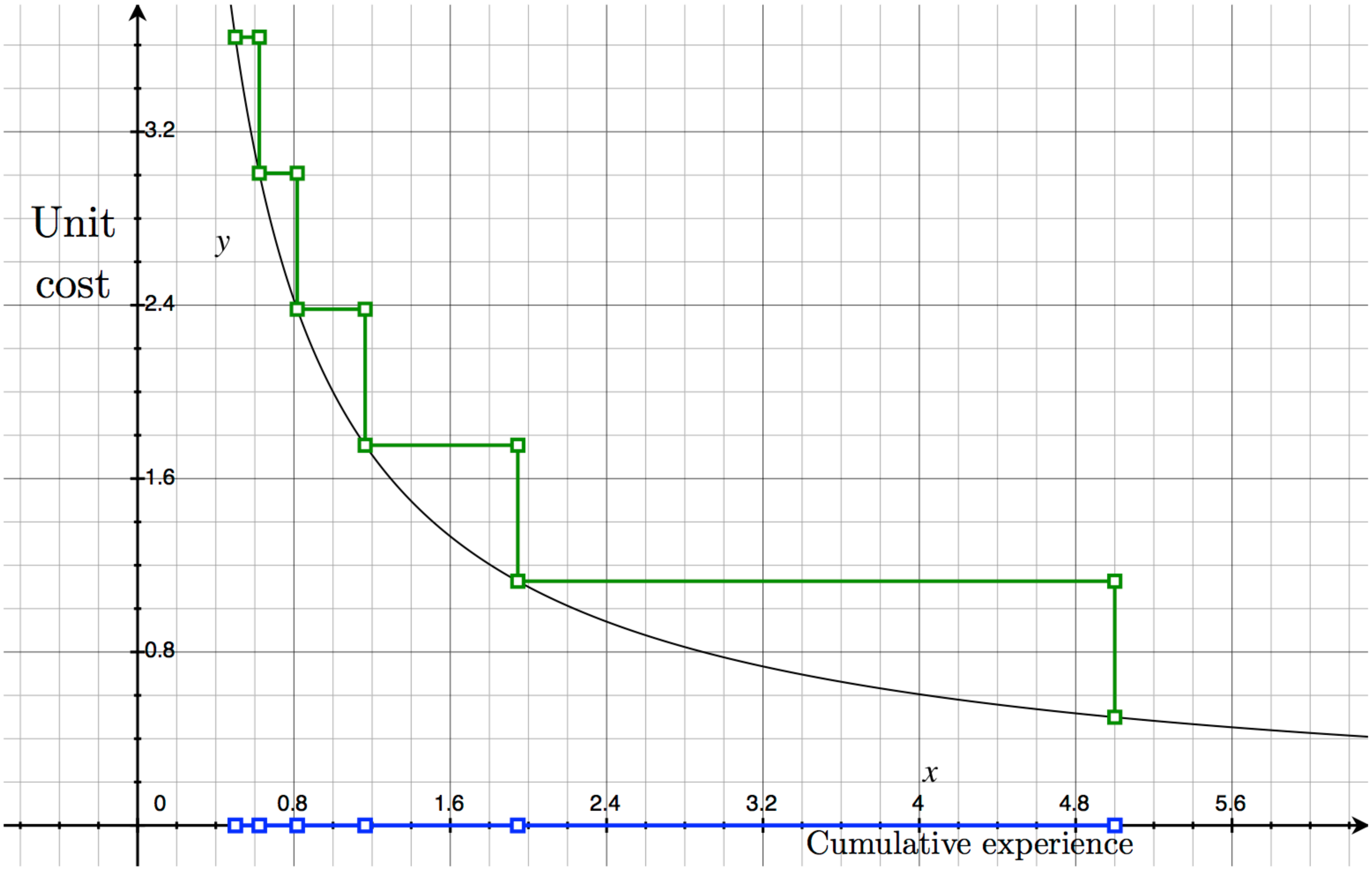
Cumulative Cost

$$F(x) = \frac{1}{1-b}x^{1-b}$$

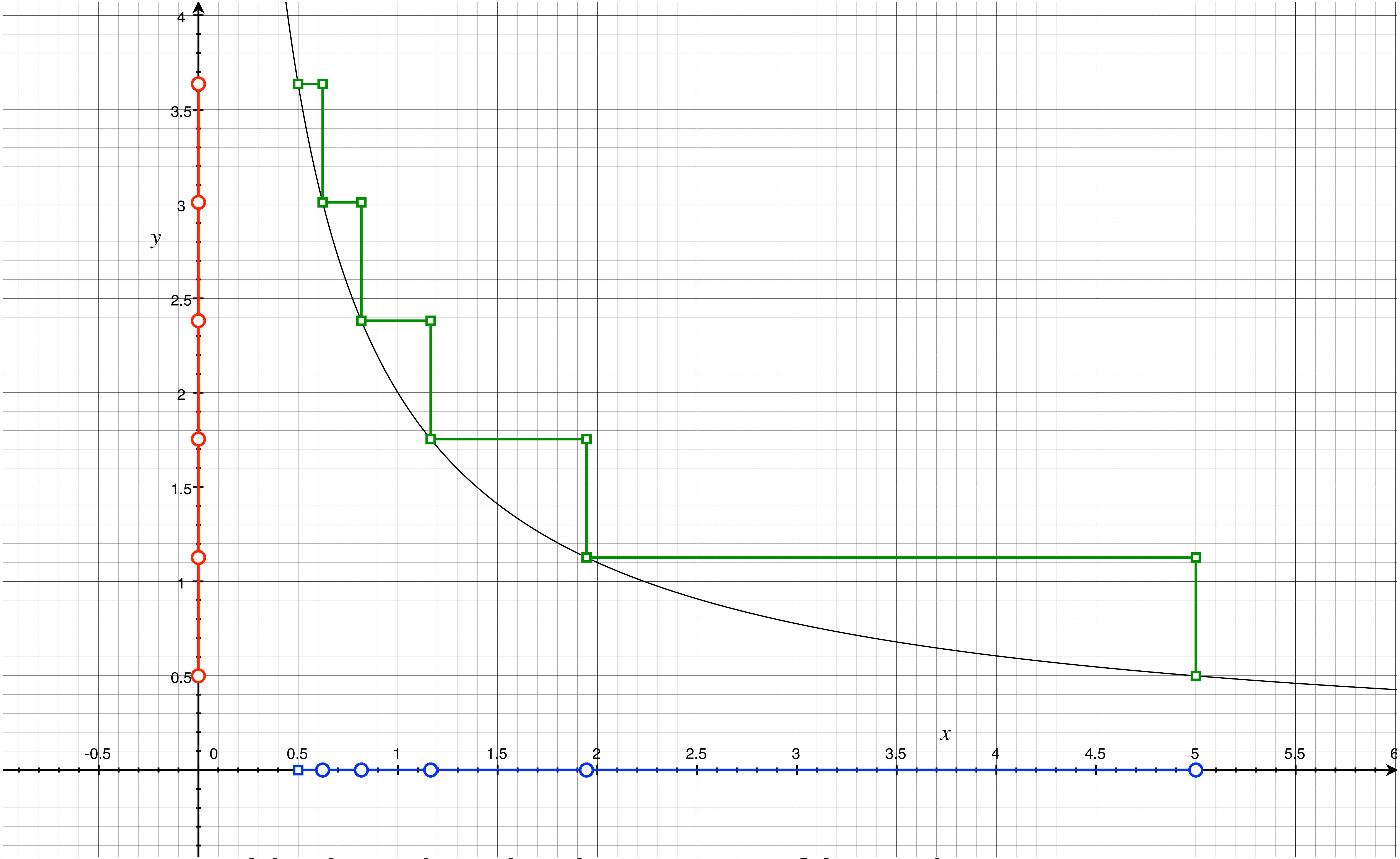
$$F(x) = \int f(x)\,dx$$

Unit Cost

$$f(x) = x^{-b}$$

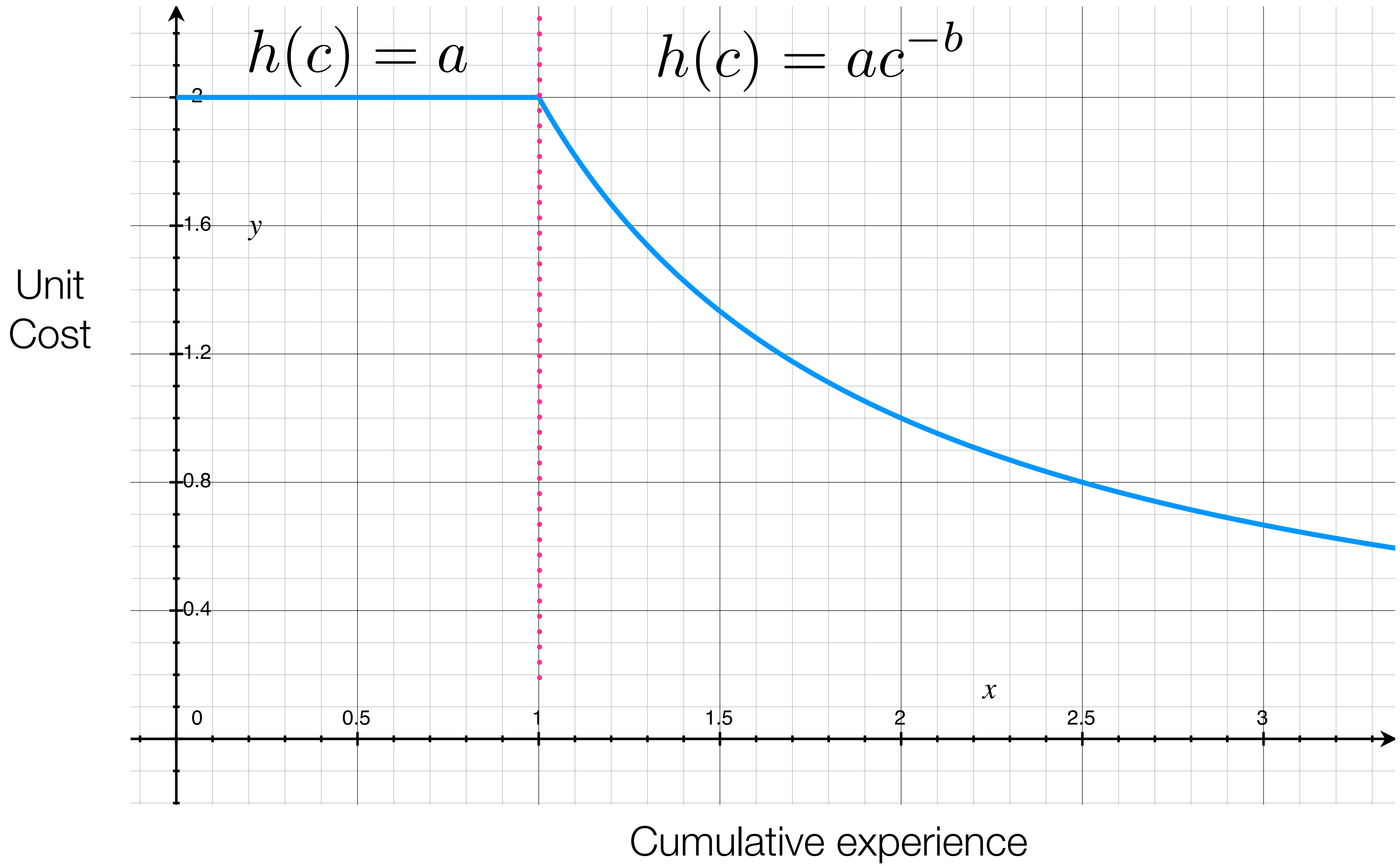


Step function approximation to unit cost curve



Notice that in the case of learning curves, the vertical spacing is **uniform**.

The unsure learning curve



Defining the problem

What is the optimal choice of *breakpoints* or *knots* within an interval $[x_{\min}, x_{\max}]$ for constructing a piecewise-linear approximation to any given function?

For a **given fixed** *N,* how do we chose the *N* knots

$$x_{\min} < t_1 \leq t_2 \cdots \leq t_N < x_{\max}$$

that achieves a piecewise-linear interpolation that is 'as close as possible'?

CSIRO

An excellent guide to **linear** approximation

de Boor, C., *Good approximation by splines with variable knots*, ISNM Vol.21, Spline functions and Approximation Theory, Birkhauser Verlag, Basel, **1973**, 57–72.

We can best approximate a function *f* by linear pieces if we choose the *N* knots

$$x_{\min} = t_0 < t_1 \le t_2 \cdots \le t_N < t_{N+1} = x_{\max}$$

so as to make each integral

$$\int_{t_i}^{t_{i+1}} \sqrt{|f''(x)|dx}$$

(approximately) the same for each *i*.

An excellent guide to **step** approximation

de Boor, C., *Good approximation by splines with variable knots*, ISNM Vol.21, Spline functions and Approximation Theory, Birkhauser Verlag, Basel, **1973**, 57–72.

We can best approximate a function *f* by linear pieces if we choose the *N* knots

$$x_{\min} = t_0 < t_1 \le t_2 \cdots \le t_N < t_{N+1} = x_{\max}$$

so as to make each integral

$$\int_{t_i}^{t_{i+1}} |f^{'}(x)| dx$$

(approximately) the same for each *i*.

Solve this equation

$$x_{\min} = t_0 < t_1 \leq t_2 \cdots \leq t_N < t_{N+1} = x_{\max}$$

de Boor, C., *Good approximation by splines with variable knots*, ISNM Vol.21, Spline functions and Approximation Theory, Birkhauser Verlag, Basel, **1973**, 57–72.

For best-approximating step functions:

$$\int_{t_i}^{t_{i+1}} |f'(x)| \, \mathrm{d}x = \frac{1}{N+1} \int_{t_0}^{t_{N+1}} |f'(x)| \, \mathrm{d}x$$

For best-approximating piecewise linear functions:

$$\int_{t_i}^{t_{i+1}} \sqrt{|f''(x)|} \, \mathrm{d}x = \frac{1}{N+1} \int_{t_0}^{t_{N+1}} \sqrt{|f''(x)|} \, \mathrm{d}x$$

Discrete data

On a discrete data set

$$(x_0, f(x_0)), (x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_M, f(x_M))$$

with the same spacing between points of size

$$h = x_{i+1} - x_i$$

we can to use a version of the second derivative called the second-order central difference:

$$f''(x_i) \approx \frac{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}{h^2}$$

Discrete data

On a discrete data set

$$(x_0, f(x_0)), (x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_M, f(x_M))$$

with the same spacing between points of size

$$h = x_{i+1} - x_i$$

we can to use a version of the integral called a sum (!)

$$\int_a^b g(s)\mathrm{d}s = \sum_{i=0}^{M} g(x_i)h$$

Putting it together

On a discrete data set

$$(x_0, f(x_0)), (x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_M, f(x_M))$$

with the same spacing between points of size

$$h = x_{i+1} - x_i$$

choose the knots

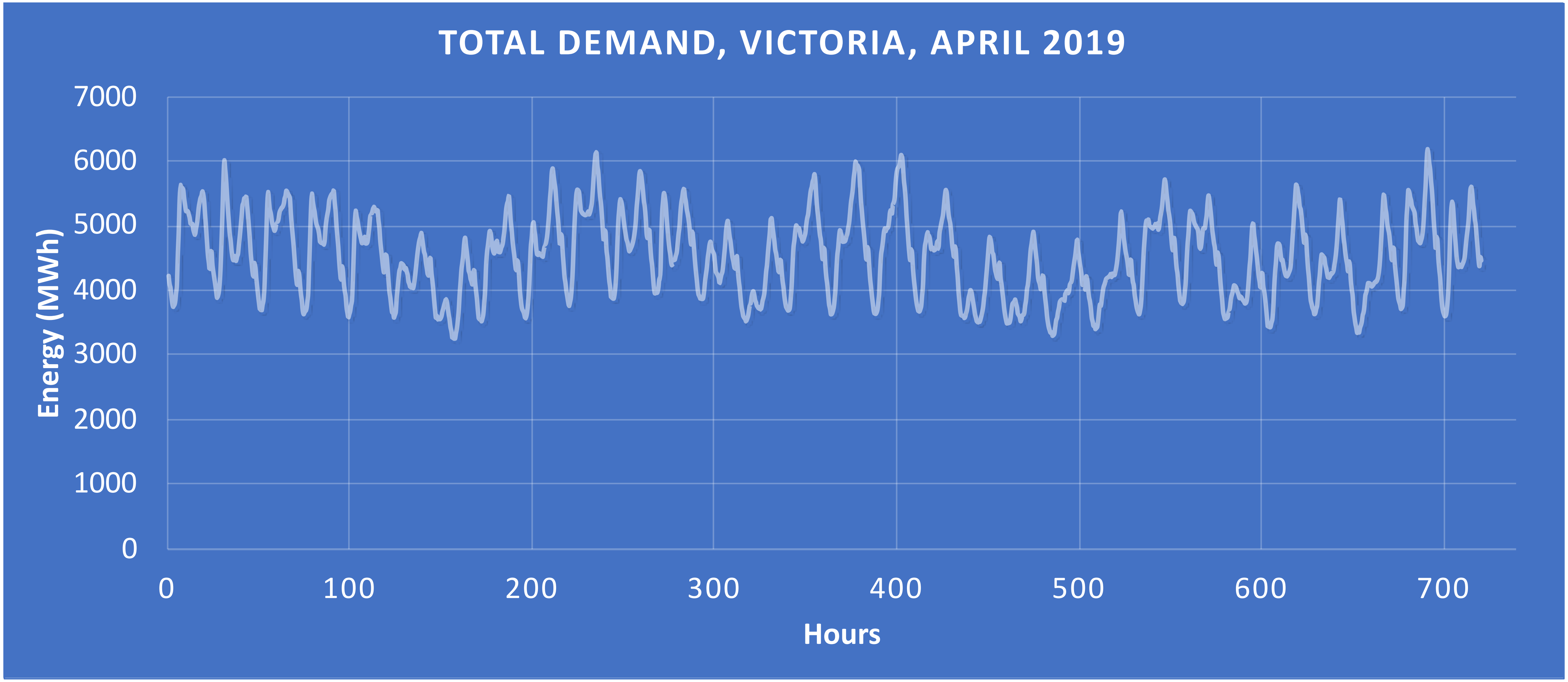
$$x_{\min} = t_0 < t_1 \leq t_2 \cdots \leq t_N < t_{N+1} = x_{\max}$$

so that **for every index** i the value

$$\int_{t_i}^{t_{i+1}} \sqrt{|f''(x)|}\,\mathrm{d}x \approx \sum \sqrt{f(x_{i+1}) - 2f(x_i) + f(x_{i-1})}$$

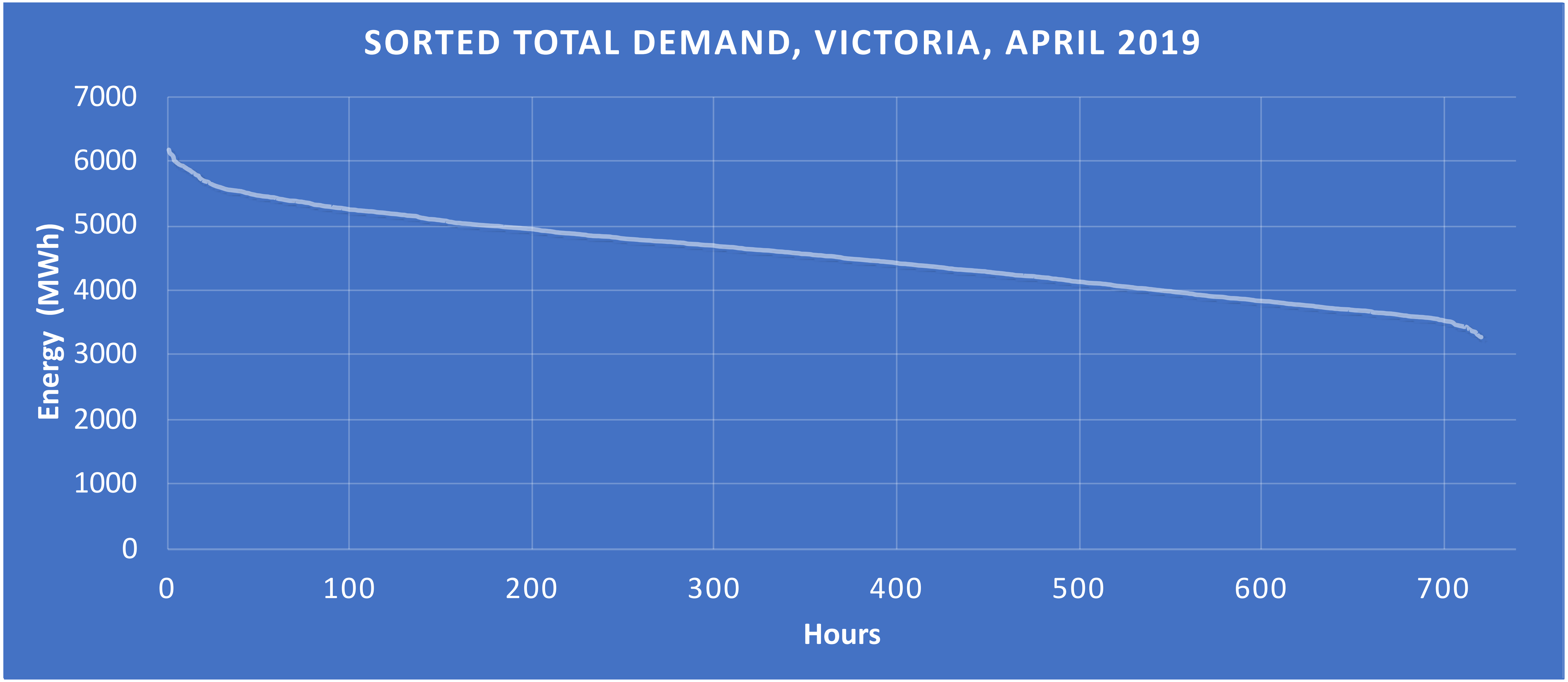
is the same, where the sum is taken over a subset of the data points $\{x_i\}$ between t_i and t_{i+1}.

Electricity demand

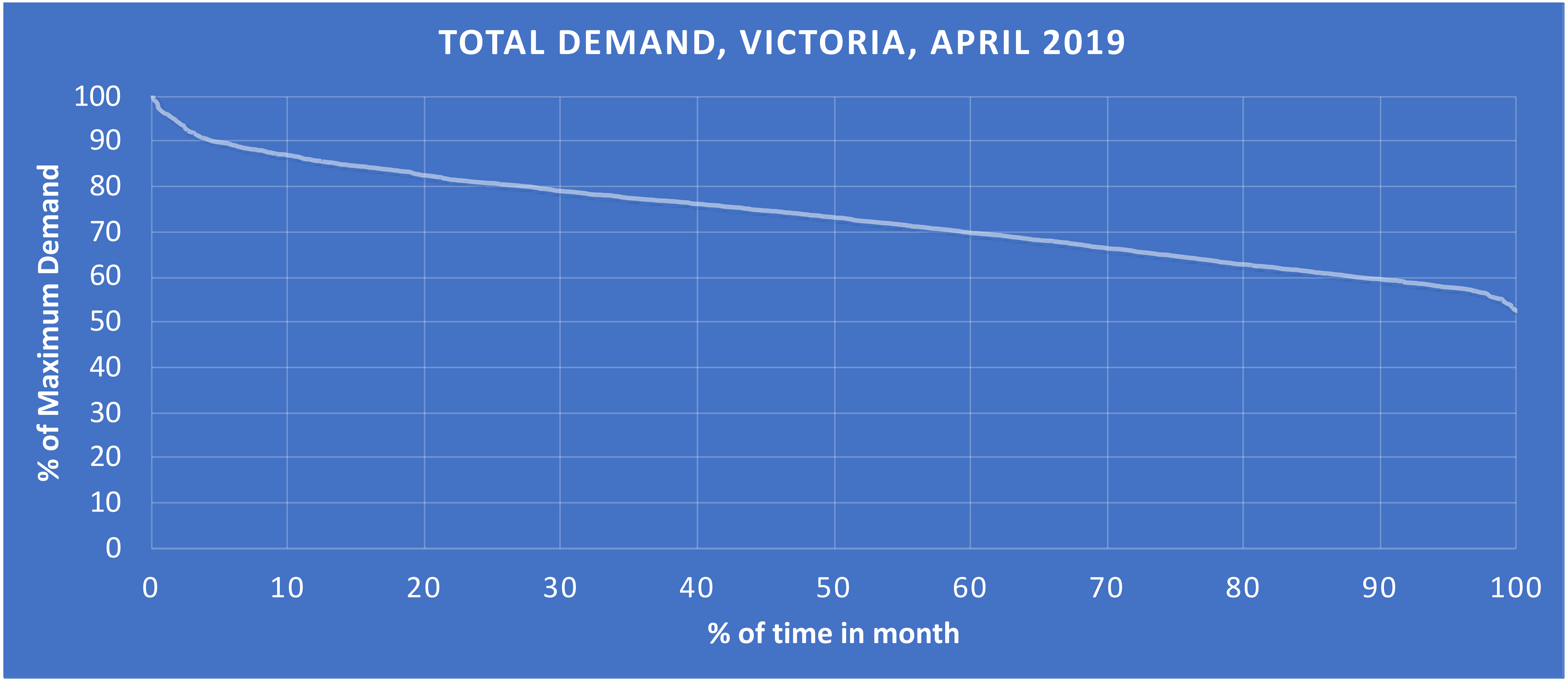


Source:
https://www.aemo.com.au/energy-systems/electricity/national-electricity-market-nem/data-nem

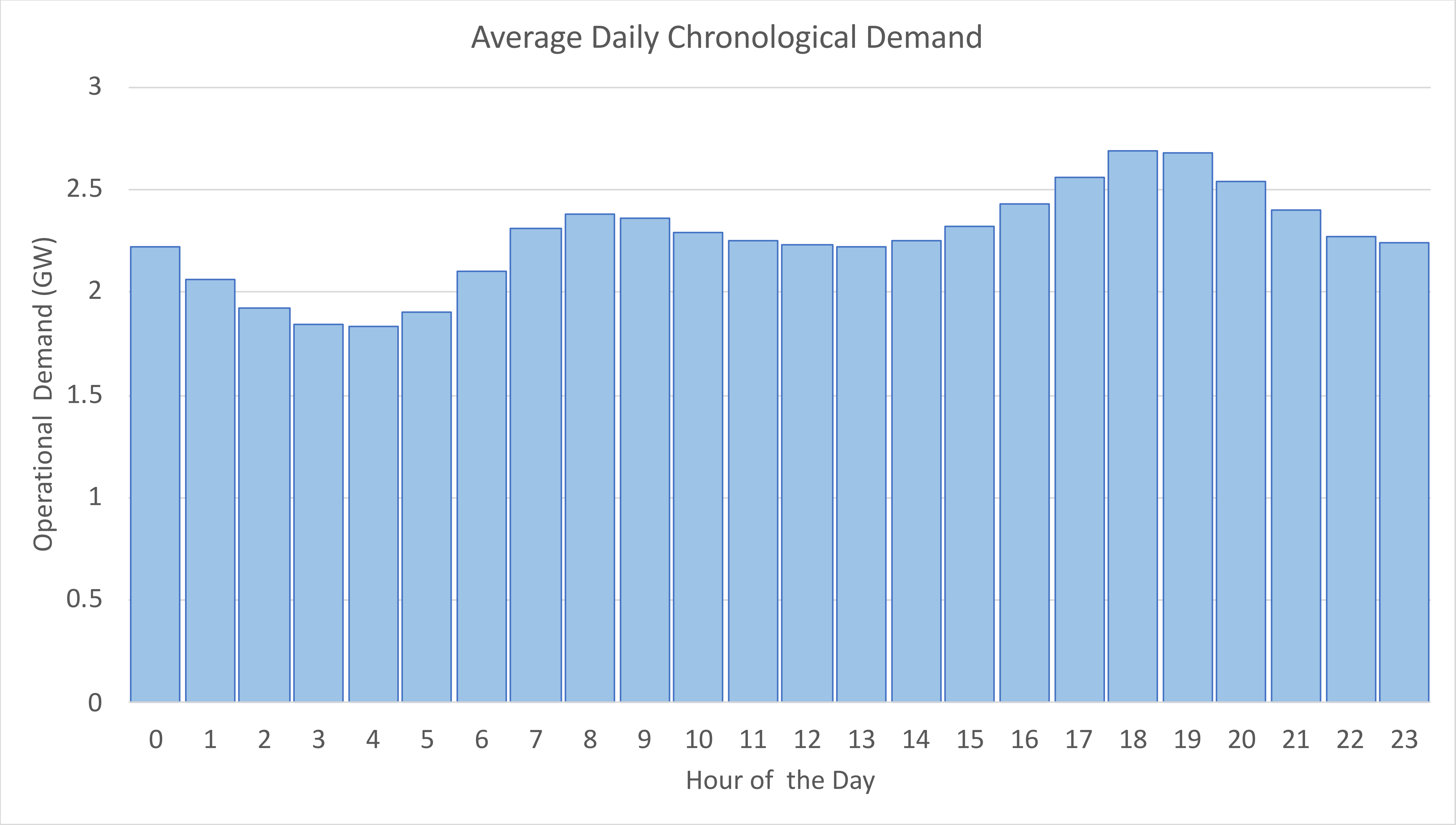
Load Duration Curves (LDCs)



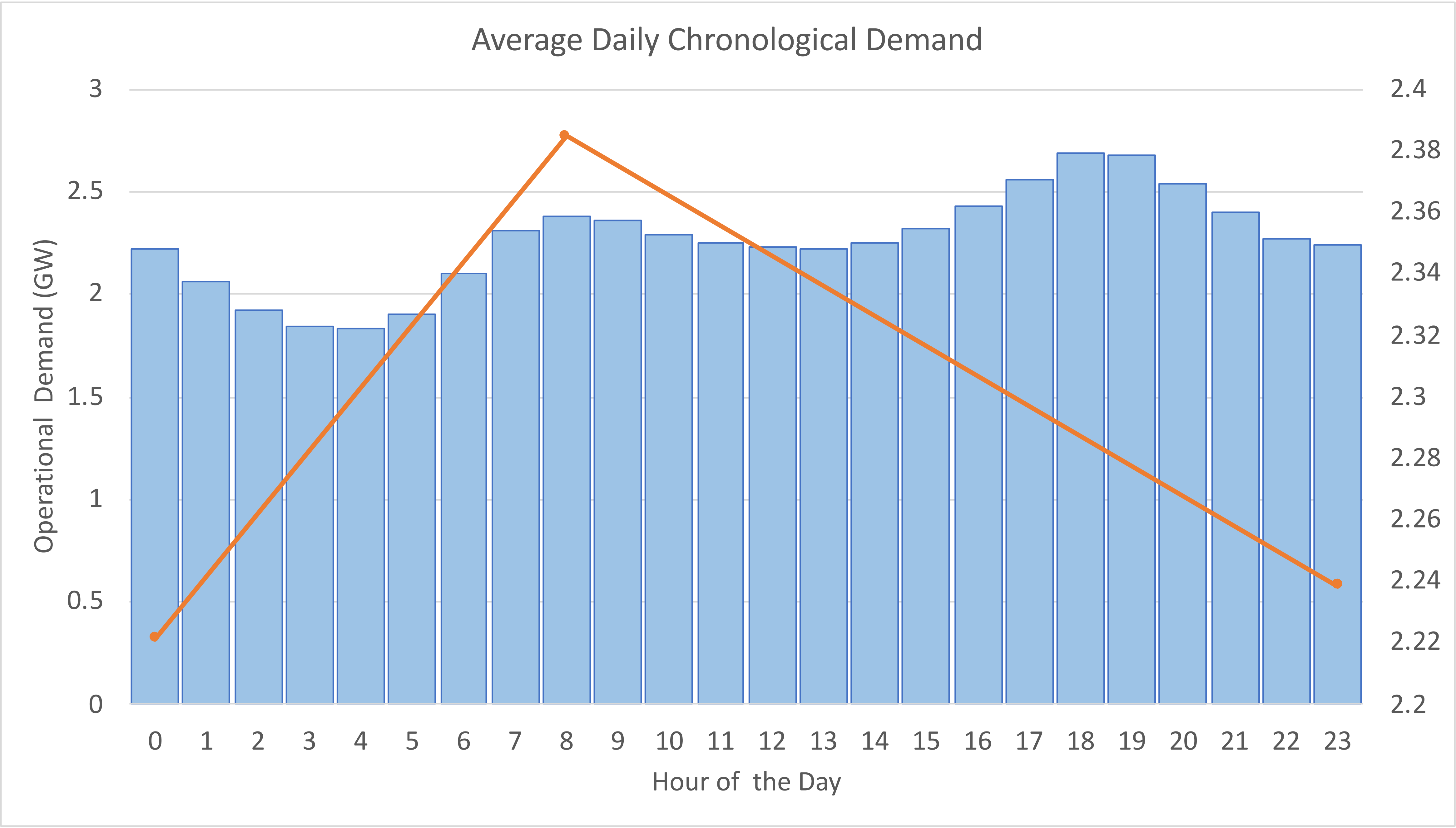
Load Duration Curves (LDCs)



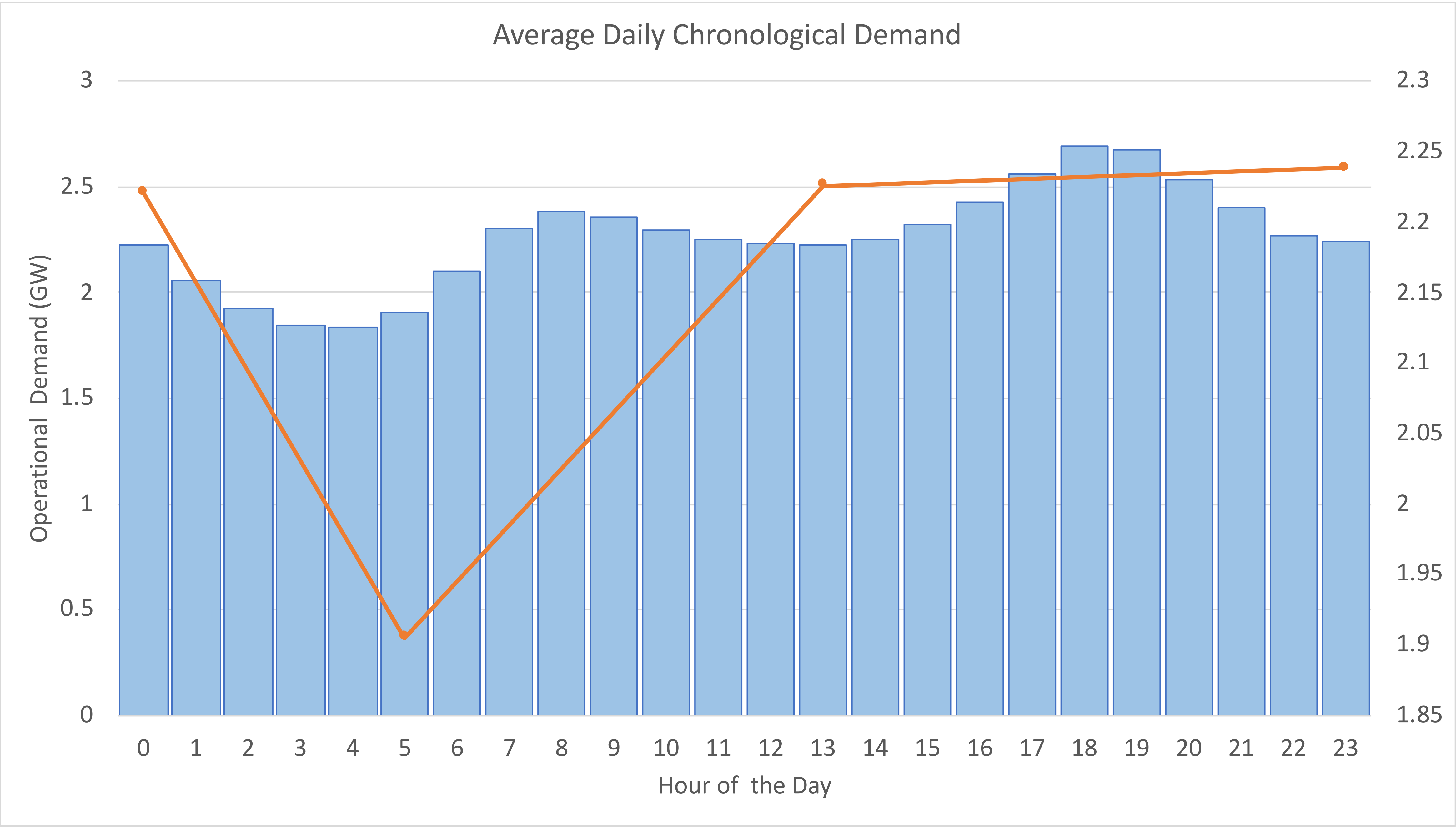
Average Daily Demand (VIC, April 2019, half-hour mean)



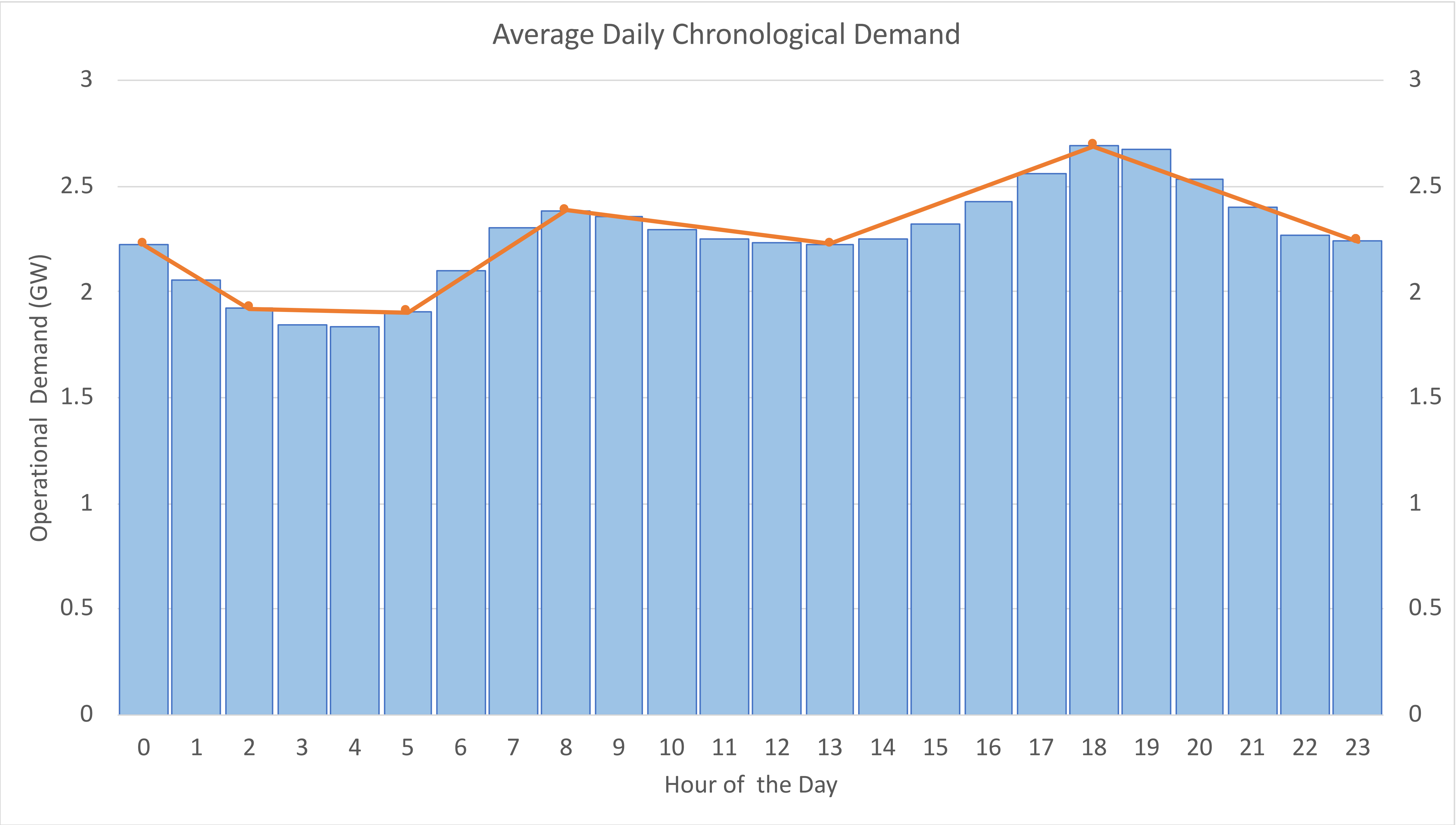
Demonstration: N=1



Demonstration: N=2

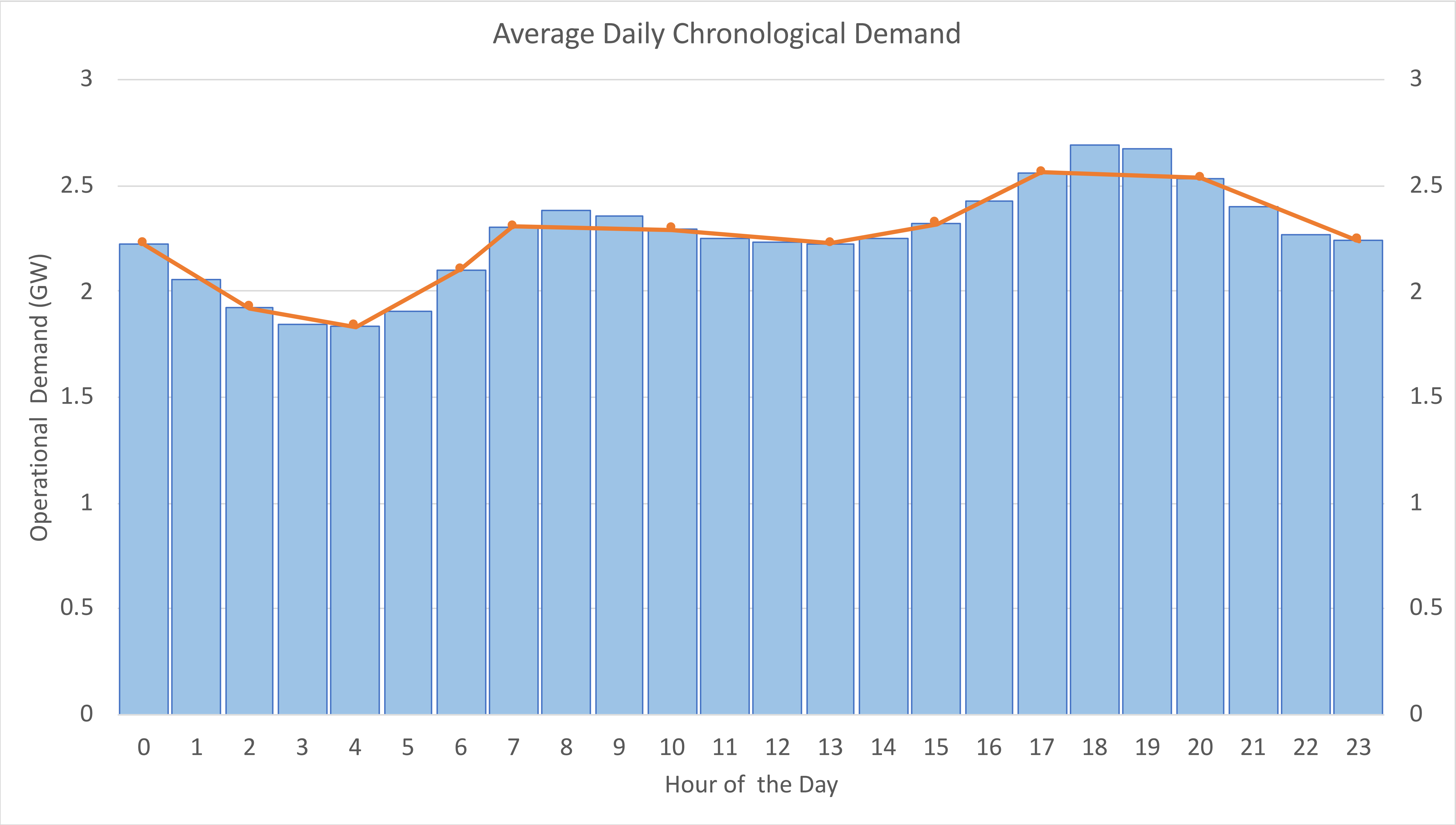


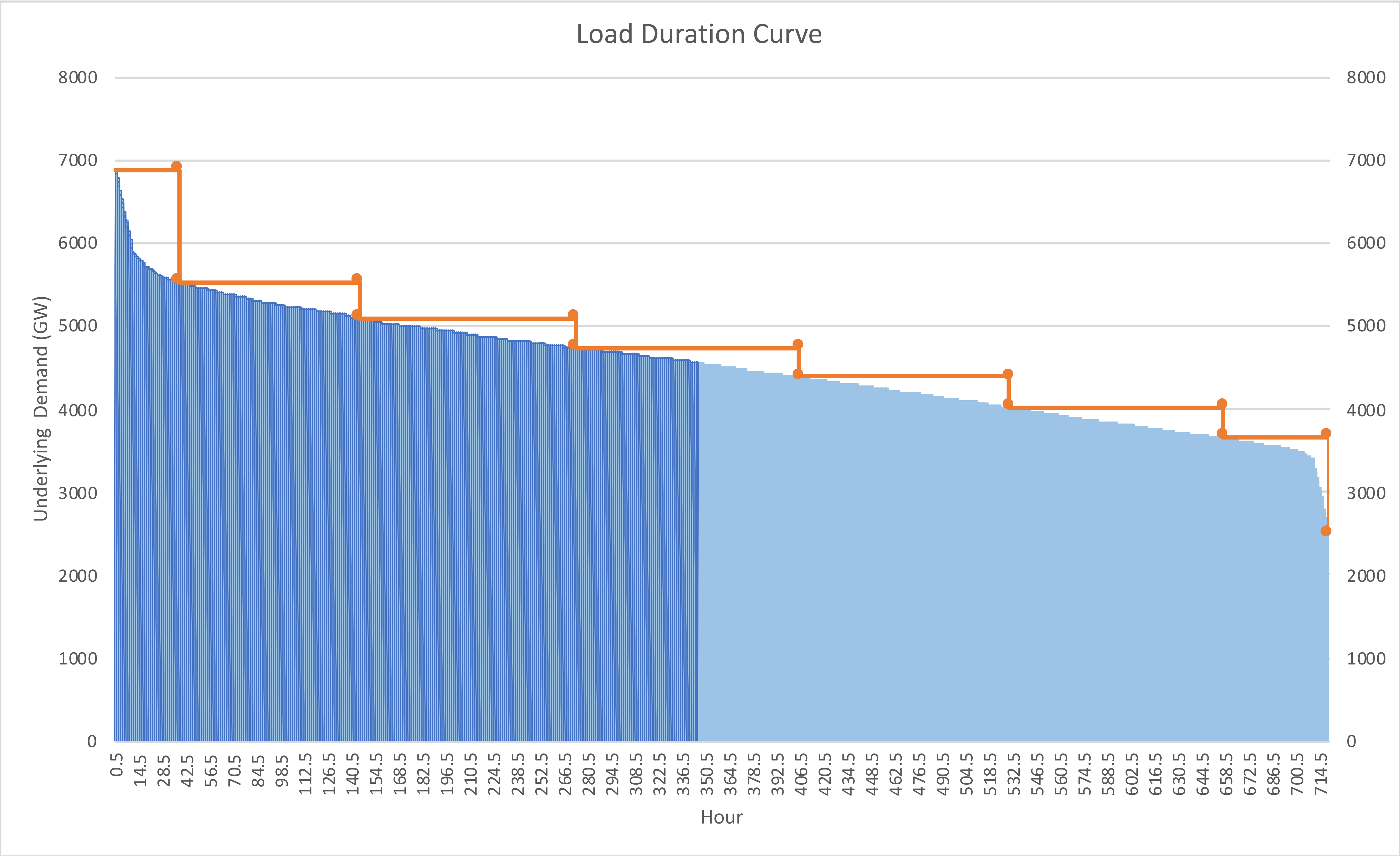
Demonstration: N=5



Demonstration: N=10

See LDC_plots_AAMT.xlsx





Load Duration Curve
Underlying Demand (GW)
Hour
0
1000
2000
3000
4000
5000
6000
7000
8000
0.5
14.5
28.5
42.5
56.5
70.5
84.5
98.5
112.5
126.5
140.5
154.5
168.5
182.5
196.5
210.5
224.5
238.5
252.5
266.5
280.5
294.5
308.5
322.5
336.5
350.5
364.5
378.5
392.5
406.5
420.5
434.5
448.5
462.5
476.5
490.5
504.5
518.5
532.5
546.5
560.5
574.5
588.5
602.5
616.5
630.5
644.5
658.5
672.5
686.5
700.5
714.5

Thank you

Energy Economic Modelling Team

Energy Systems

James Foster

t +61 2 4960 6055
e james.foster@csiro.au
w www.csiro.au