Evaluating Forecasting Performance in the Context of Process-Level Decisions:

Methods, Computation Platform, and Studies in Residential Electricity Demand Estimation

Submitted in partial fulfillment of the requirements for
the degree of
Doctor of Philosophy
in
Engineering and Public Policy

Richard A. Huntsinger

B.S., Computer Science, California State University, Chico
M.S., Computer Science, California State University, Chico
M.B.A., University of California, Berkeley
M.B.A., Columbia University

Carnegie Mellon University
Pittsburgh, Pennsylvania

May 2017

For Nana

ABSTRACT

"To measure is to know."

– William Thomson, First Baron Kelvin

This dissertation explores how decisions about the forecasting process can affect the evaluation of forecasting performance, in general and in the domain of residential electricity demand estimation. Decisions of interest include those around data sourcing, sampling, clustering, temporal magnification, algorithm selection, testing approach, evaluation metrics, and others.

Models of the forecasting process and analysis methods are formulated in terms of a three-tier decision taxonomy, by which decision effects are exposed through systematic enumeration of the techniques resulting from those decisions. A computation platform based on the models is implemented to compute and visualize the effects. The methods and computation platform are first demonstrated by applying them to 3,003 benchmark datasets to investigate various decisions, including those that could impact the relationship between data entropy and forecastability. Then, they are used to study over 10,624 week-ahead and day-ahead residential electricity demand forecasting techniques, utilizing fine-resolution electricity usage data collected over 18 months on groups of 782 and 223 households by real smart electric grids in Ireland and Australia, respectively.

The main finding from this research is that forecasting performance is highly sensitive to the interaction effects of many decisions. Sampling is found to be an especially effective data strategy, clustering not so, temporal magnification mixed. Other relationships between certain decisions and performance are surfaced, too. While these findings are empirical and specific to one practically scoped investigation, they are potentially generalizable, with implications for residential electricity demand estimation, smart electric grid design, and electricity policy.

DISSERTATION IN THREE SENTENCES

"If you can't explain it simply, you don't understand it well enough."

– Albert Einstein

We formulate methods and develop software to explore various ways of evaluating forecasting performance in the context of process-level decisions. We then conduct about 800,000 experiments using data from smart electric grids in Ireland and Australia, and from other sources. We find several potentially generalizable relationships between decisions and performance, the primary one being that performance depends on the interaction effects of many decisions, not just a few.

ACKNOWLEDGEMENTS

"I get by with a little help from my friends."

– John Lennon and Paul McCartney

The candidate very much thanks the dissertation committee for valuable guidance on this research:

Paul Fischbeck (chair)	Department of Engineering & Public Policy, Carnegie Mellon University. Department of Social and Decision Sciences, Carnegie Mellon University. Scott Institute for Energy Innovation, Carnegie Mellon University.
Eduard Hovy	School of Computer Science, Carnegie Mellon University. Command, Control, and Interoperability Center for Advanced Data Analysis, Rutgers University.
Curt Powley	Department of Computer Science, Hawaii Pacific University.
Mitchell Small	Department of Engineering & Public Policy, Carnegie Mellon University. Department of Civil and Environmental Engineering, Carnegie Mellon University. Scott Institute for Energy Innovation, Carnegie Mellon University.

This research was possible only with assistance and encouragement from Hadi Amini, Jeffrey Anderson, Michael Ford, Justin Frye, ... and mostly from my dear wife, Diane.

This research was self-supported.

TABLE OF CONTENTS

1 INTRODUCTION

"The beginning is the most important part of the work."

– Plato

Policies to address future electricity demand are informed by forecasting techniques involving statistical and more recently machine learning models derived from historical electricity usage, weather, and other data. Now, the increasing availability of large-scale, stratified, fine-resolution electricity usage data from smart electric grids that aligns regionally and temporally with weather data is enabling forecasting research based on detailed knowledge of actual electricity consumer behavior.

In this dissertation, we are specifically interested in leveraging such smart electric grid data to explore how forecasting process decisions can affect the evaluation of forecasting performance, and how sensitive such evaluation is to various data strategies and data sources, especially in the domain of residential electricity demand estimation, and to better understand the implications for smart electric grid design and related electricity policies.

This dissertation is organized into 7 chapters and appendices, as follows:

Chapter 1: Introduction We introduce the research and lay out the dissertation organization.

Chapter 2: Models, Methods, and Computation Platform We introduce domain-agnostic models of the forecasting process and associated analysis methods in terms of a three-tier decision taxonomy, with which we can explore the effects of forecasting process decisions on the evaluation of forecasting performance by systematically enumerating techniques to construct, score, and rank forecast distributions. To practically realize the methods, we implement a corresponding computation platform to calculate and visualize these effects. We

demonstrate the methods and computation platform by exploring effects around the M3 Forecasting Competition datasets.

Chapter 3: Decisions and Data Entropy as a Predictor of Forecastability We analyze the effects of forecasting process decisions on the evaluation of forecasting performance, and on the relationship between entropy and forecastability, using the M3 Forecasting Competition datasets. We compare our results to those of some benchmark studies and propose quantitative relationships between entropy and forecastability.

Chapter 4: Technique Decisions and Residential Electricity Demand Estimation We analyze the effects of forecasting process decisions on the evaluation of forecasting performance when forecasting week-ahead and day-ahead residential electricity demand using real smart electric grid data from Ireland. We compare our results to those of some benchmark studies.

Chapter 5: Data Strategy Decisions and Residential Electricity Demand Estimation We further analyze decision effects on residential electricity demand forecasting, leveraging earlier results and focusing on decisions for three specific data strategies: sampling, clustering, and temporal magnification.

Chapter 6: Robustness of Decision Effects and Residential Electricity Demand Estimation We further analyze decision effects on residential electricity demand forecasting, leveraging earlier results and focusing on how the effects may generalize across multiple locations, as represented by multiple data sources, adding real smart electric grid data from Australia.

Chapter 7: Conclusion We summarize our insights and suggest future research.

Appendices We provide primers on a few topics important to this research to supplement our discussions in the main body.

To enhance readability, discussions are presented mostly uninterrupted by in-line tables and data visualizations, and rather reference table and data visualization compilations at the ends of the chapters. Reference articles are referenced by number throughout and organized by

topic at the end of the document. Certain specifications and other list-like portions of discussions are set out in gray boxes. Key insights from our results are set out in double-border gray boxes. Per popular usage, the term “data” is used in both singular and plural senses, distinguishable by context. Similarly, the term “series” is used in both singular and plural senses, distinguishable by context.

2 MODELS, METHODS, AND COMPUTATION PLATFORM

"A place for everything, everything in its place."

– Benjamin Franklin

2.1 Research Questions

In this chapter, we address the following research questions:

- What is the forecasting process and what decisions are involved?
- How do forecasting process decisions affect forecasting performance?
- How can forecasting performance be compared across studies?
- How can data be sourced and structured for forecasting performance analysis?

2.2 Literature Review

Forecasting studies reported in the academic and professional literature over several decades have treated the forecasting process with more or less formality depending on the study focus. Various meta-studies have attempted to generalize the forecasting process in terms of decision environment, forecast object, forecast statement, forecast horizon, information set, complexity, or other dimensions. [3,26]

Algorithm selection, indeed algorithm invention, and associated hyper-parameter selection account for a large portion, if not most, of the recent research on forecasting techniques (Table 2-3, Table 2-4, Table 2-5). [14,17,25,37,40,74,75] By algorithm, we mean that part of the forecasting process that constructs an engine that, in combination with other elements of a technique, can construct forecasts – an algorithm directly constructs an engine and thereby indirectly constructs forecasts. By hyper-parameters, we mean those algorithm parameters outside of training data that influence engine construction, i.e., the parameters used to

configure algorithms. In many studies, some algorithm details are called out separately as transformation operators, which pre-process the training data or inputs (Table 2-6). [2]

Two broad classes of algorithms have been studied extensively: statistical and machine learning. Statistical algorithms take time series as inputs and output point forecasts. They have been applied to general forecasting since at least the 1950s. [2,76] In contrast, machine learning algorithms train on cross-sectional data to produce engines, which in turn take other cross-sectional data as inputs and output point forecasts. Machine learning came into vogue for general forecasting by the 1980s. [42,76]

More recently, research on extension rules has accelerated (Table 2-7). [43,44,45,46,47,48] By extension rule, we mean the way in which point forecasts from engines are compiled into time series forecasts. A "direct" extension rule assumes that an engine makes a point forecast, so a set of engines is required to produce a forecast covering a period, one engine for each time step in the time series. A "recursive" extension rule assumes that an engine makes a point forecast one time step farther into the future, so the engine must be applied from successive time steps to produce a forecast covering a period. In a recursive extension rule, each new point forecast is determined based partly on earlier point forecasts.

Different training and testing rules can, in general, result in different forecasts. By training and testing rule, we mean how data is apportioned to training and testing activities. We find limited research on training and testing rules (Table 2-8). [49,50,51,52,53] Rather, forecasting studies often adopt a conventional holdout training and testing rule where forecasts are evaluated with respect to the later part of available actual historical data that is treated as a perfect forecast. The proportion of data reserved for testing may be based on rules of thumb, often 20% or 33%. [13,54] Other research describes time series cross-validation training and testing rules. [49,50,51,52,53,54]

An ever-increasing array of proposals for new ways to score forecast performance continues to inhere in the literature, with each metric purporting its suitability for various general or special cases (Table 2-9). [2,56,57,58,61,63,66] Depending on the application, various studies have

used metrics from one or more classes: correlation metrics, relative error metrics, absolute error metrics, or penalty functions, which are essentially customizable metrics that value errors per their economic impact.
[18,80,88,90,91,93,95,96,97,98,99,101,102,105,107,108,109,110,112,113,115,116,117,118]

Among the many forecasting studies are analyses of forecasting competitions. Notable general forecasting competitions include the Makridakis competitions (M, M2, M3, M4) and the Neural Networks Forecasting Competition (NN3). [134,135,136,141,154,156] These competitions primarily pit algorithms against each other, constraining data strategy, extension rule, training and testing rule, metrics, and other aspects of the forecasting process to ensure uniform competition environments.

We find potential application of research on relative importance, which quantifies how independent variables disproportionately account for variation in a dependent variable, to research about forecasting, but not well represented in the literature. Several measures of relative importance are in use, perhaps LMG being the most common. [144,145,146,147]

In summary, we find the following prevalent themes in forecasting research:

Forecasting Process	• There are many formalizations of the forecasting process.
Forecasting Process Evaluation	• Studies often evaluate forecasting performance in the context of one or two decisions at a time.

2.3 Research Approach

Our literature review motivates us to investigate the effects of many forecasting process decisions working in concert to better understand the impact of their interactions. Our approach is to introduce explicit and exhaustive methods to analyze forecasting performance, based on models of the forecasting process that account for process-level decisions. Specifically, we formalize notions of “data source”, plus “data strategy”, “technique”, and “metric” as a three-tier decision taxonomy, with each level corresponding to a model of the

forecasting process that addresses a certain category of explicitly identified decisions. A vector of all decisions then uniquely instantiates the model to a specific version of the overall forecasting process. Potential decision combinations, expressed as vectors, are exhaustively enumerated by cycling through a set of allowed options over which the decisions can range. The set of vectors can then be used to construct and score forecasts, map data strategies and techniques to scores, and rank data strategies and techniques accordingly – in a computationally intensive exercise (Figure 2-1).

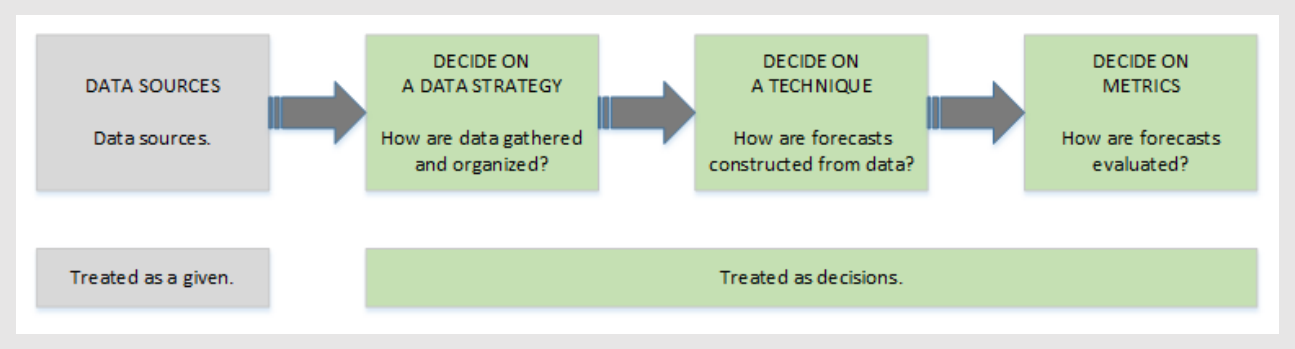


Figure 2-1: Three-tier decision taxonomy.

By data source, we mean a specific set of data with which to train for and test forecasts. By data strategy, we mean a specific way in which a forecasting process model is instantiated by decisions about how data are gathered and organized. By technique, we mean a specific way in which a forecasting process model is further instantiated by decisions about objectives, form, and training and testing. By metric, we mean an operator used to quantify forecasting performance. By forecasting performance, we mean an evaluation of the distribution of estimated forecast accuracy across the set of data strategies, techniques, and metrics.

By accounting for the effects of many forecasting process decisions working in concert rather than in isolation, our methods are not restricted to making “apples to apples” comparisons of algorithm class, metric, or other single decision, but are empowered to make potentially more insightful “fruit baskets to fruit baskets” comparisons of whole forecasting processes.

2.4 Data Sources

The data sources determine what data to use to train for and test forecasts. Our model of the forecasting process that addresses data sources is straight-forward: data sources are treated as a given and reserved for later use by a forecasting process that addresses technique decisions.

The data source includes the following.

Reference data source	A dataset on which models will be trained and forecasts evaluated.	a data file which is or can be converted to a set of equal length time series, including datetime & aggregator information
Predictor data sources	A set of datasets on which models will be trained and forecasts evaluated.	a set of data files, each of which is or can be converted to a set of equal length time series, including datetime & aggregator information

2.5 Data Strategy Decisions

Data strategy decisions determine the way in which data is gathered and organized. Specifically, they determine how data is transformed to a sample, a group of clusters, a temporal magnification, or other form, and how those transformations are sequenced.

2.5.1 Strategic Forecasting Process Model

We call the most general form of our model of the forecasting process, which exposes how data strategy decisions are handled, the strategic forecasting process model (Figure 2-2). Per the three-tier decision taxonomy, the model assumes data sources have been provided. It accounts for integration rules indicating in which ways sampling, clustering, temporal magnification, and

aggregation of data can be combined and sequenced. Each of these data transformations can be applied singly or multiply in isolation or in combination, such that 132 unique sensible integration rules (paths) are possible (Table 2-10, Table 2-11, Table 2-12).

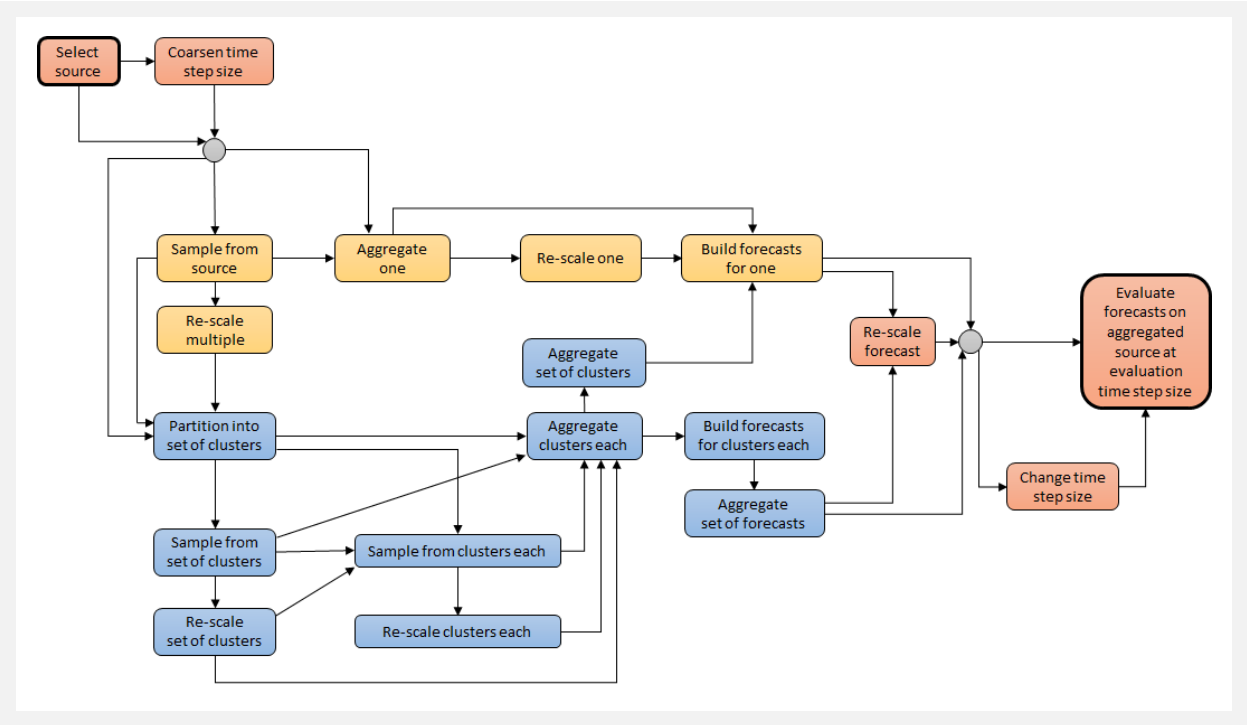


Figure 2-2: Strategic forecasting process model. Each path corresponds to a unique integration rule. *Yellow* indicates tasks that operate on non-clustered data. *Blue* indicates tasks that operate on clustered data.

2.5.2 Decision Descriptions

Data strategy decisions include the following. These decisions are typically made by the forecasting practitioner.

Sampling rule	A rule indicating how (if at all) datasets should be resampled. Rules like bootstrapping and jackknife sampling provide control over the degree of randomness in the resulting resampled datasets.	`full population` or `bootstrap v1` or `bootstrap v2` or `bootstrap v3` or `jackknife` or other rule
Sample size	A count of the number of individual series aggregated into a sample. Can correspond to, e.g., individual households up to neighborhoods, districts, or any arbitrary larger sample.	a count
Cluster similarity criteria	An operator on two series used to judge their similarity.	`correlation` or `coherence` or other operator
Cluster count	A number of clusters by which a population is partitioned.	a count
Time step size – pre-integration	A period corresponding to the time step size used to construct forecasts before they are aggregated and evaluated. It can be different than the reference series or predictor series time step sizes, such as when making hourly forecasts based on data collected semi-hourly.	a period
Integration rule	An indication of the sequence in which data are sampled, clustered, temporally magnified, aggregated, and re-scaled.	an integration rule index i in $1 \leq i \leq 132$
Other data strategy decisions …		

See more about sampling rules in Appendix A.

2.5.3 Primary Integration Rules

Four specific integration rules instantiate the strategic forecasting process model in ways convenient for studying sampling, clustering, and temporal magnification separately. We call these the primary integration rules (Figure 2-3).

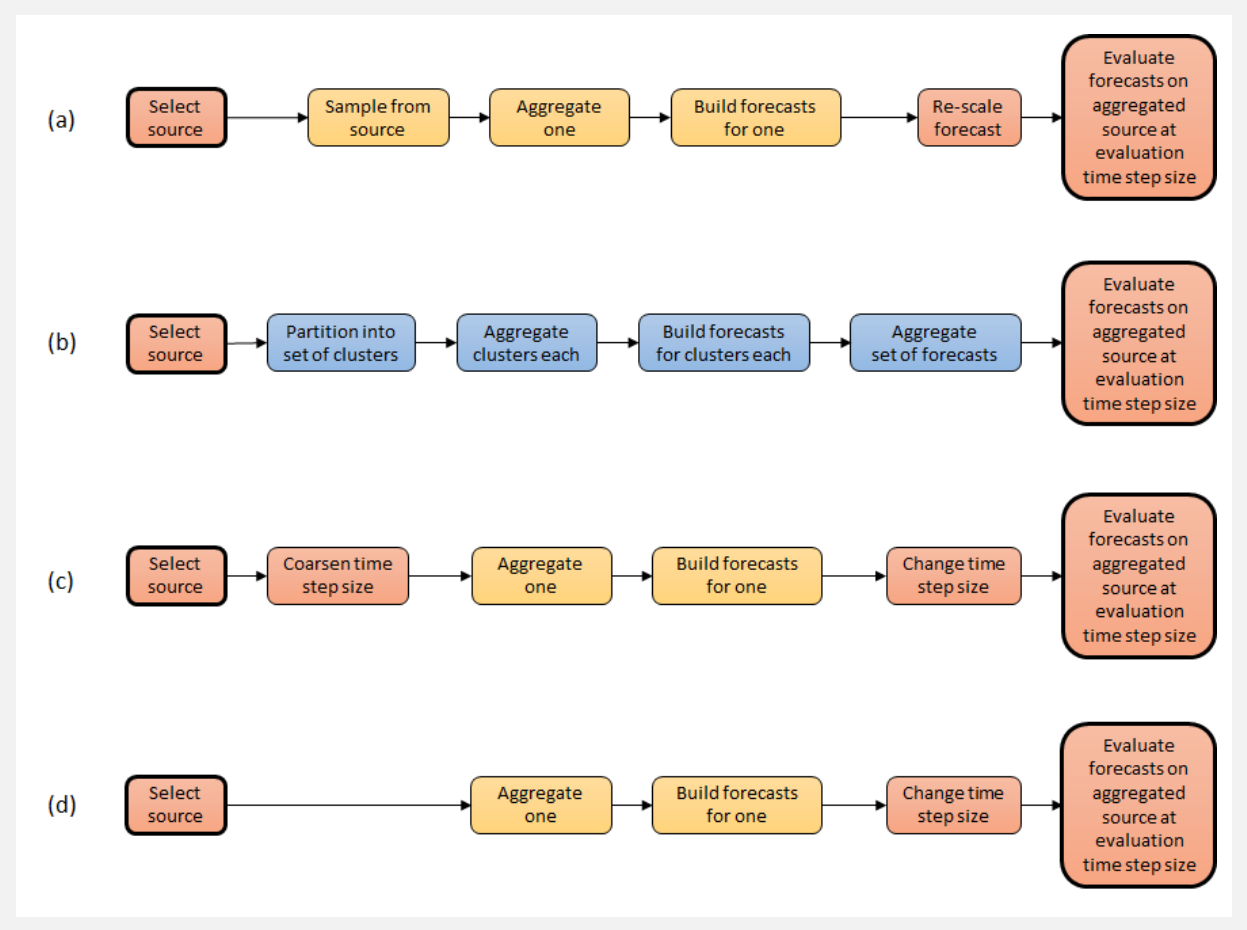


Figure 2-3: Strategic forecasting process instantiations per the four primary integration rules. *Yellow* indicates tasks that operate on non-clustered data. *Blue* indicates tasks that operate on clustered data.

Primary Integration Rule: Forecast Based on a Sample

To forecast based on a sample, first select a data source to serve as the reference series. From this population of series, take a sample of series per a sampling rule and aggregate them to produce a single sample aggregate series. Next, informed by a technique, construct a single sample aggregate series forecast. Then, re-scale this forecast in proportion to the whole population, such that the re-scaled forecast becomes a population series forecast, which can be evaluated against the actual population series.

Summary: Sample source, re-scale after forecasting.

Primary Integration Rule: Forecast Based on a Group of Clusters

To forecast based on a group of clusters, select a data source and partition it into a set of clusters per some measure of similarity. For each cluster, aggregate its member series to produce a single cluster aggregate series reflecting the aggregate levels for the cluster, and so get a set of cluster aggregate series forecasts. Next, for each cluster and a technique, construct a single cluster aggregate series forecast. Then, aggregate the cluster aggregate series, and so get a population series forecast, which can be evaluated against the actual population series.

Summary: Partition source into clusters, forecast each cluster, aggregate forecasts.

Primary Integration Rule: Forecast Based on Coarsening Temporal Magnification

To forecast based on coarsening temporal magnification, select a data source and coarsen the time step size of each series, and so get a set comprising the same number of series that the data source does, but with each series shorter in proportion to how much it is coarsened (i.e., each series comprises fewer time steps, with each time step representing a longer duration). From this population of series, aggregate to produce a single population coarse-resolution aggregate series. Next, informed by a technique, construct a single population coarse-

resolution aggregate series forecast. Then, refine the time step size of the forecast, which can be evaluated against the actual population fine-resolution series.

Summary: Forecast on coarse time step size, evaluate on fine time step size.

Primary Integration Rule: Forecast Based on Refining Temporal Magnification

To forecast based on refining temporal magnification, do the reverse of forecasting by coarsening, initially leaving the data source at a fine time step size, and later coarsen the forecasts to prepare them for evaluation against the actual population coarse-resolution series.

Summary: Forecast on fine time step size, evaluate on coarse time step size.

2.6 Technique Decisions

Technique decisions determine how forecasting engines and forecasts themselves are constructed.

2.6.1 Basic Forecasting Process Model

We call the part of our model of the forecasting process that exposes how technique decisions are handled the basic forecasting process model. Per the three-tier decision taxonomy, this part assumes data sources and data strategies have been provided and are handled at higher levels, and contents itself as a sub-process nested within the strategic forecasting process model.

In the basic forecasting process model, forecasting is accomplished in two phases: (1) an algorithm is used to construct a forecasting engine or engines, and (2) the forecasting engines are then used to ultimately construct a forecast (Figure 2-4).

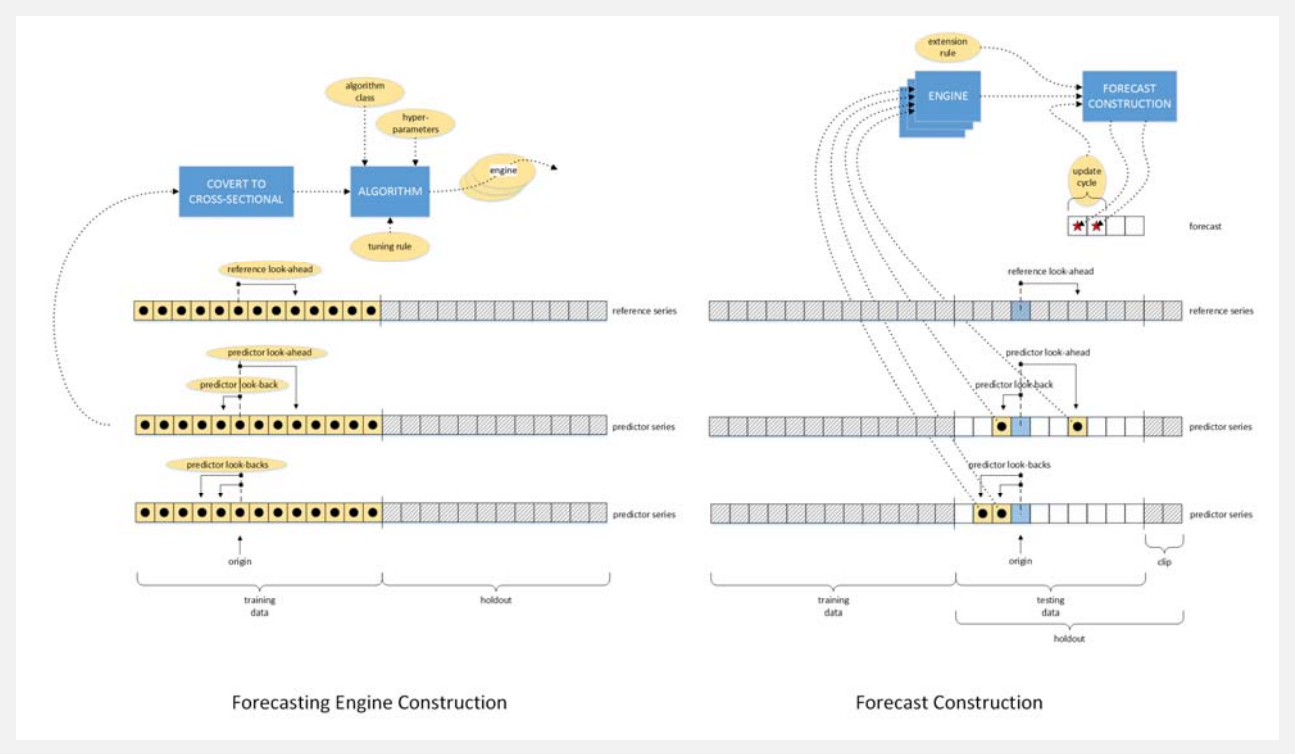


Figure 2-4: Basic forecasting process model. Part of the strategic forecasting process model.

Forecasting Engine Construction Phase

In the forecasting engine construction phase, forecasting objectives are established and time series data is acquired to train forecasting engines and test forecasts. We call the data that compares to the forecasting engine output the reference series. We call the data used as inputs to the forecasting engine the predictor series. The reference series is often also a predictor series, but it need not be. The reference and predictor series from their various sources may span different periods and be captured at different time step sizes, so they must be compiled into a data ensemble, such that start time, stop time, and time step size are aligned.

A portion of the data ensemble is reserved for training and another portion for testing. The training data could precede, follow, or be interspersed with the testing data. We call the temporal relationship between training and testing data the holdout rule. We call the portion

of data not reserved for training the holdout. We call any portion of the holdout not used for testing the clip, which can be used to conveniently avoid testing on any particularly anomalous portion of data.

A reference look-ahead, established earlier as a forecasting objective, and predictor look-backs and look-aheads instruct an algorithm how far ahead or behind an origin time step to forecast a data point or how to find predictor data points, respectively. Look-backs and look-aheads are, in general, different for each predictor series. Predictor look-aheads are functionally equivalent to negative look-backs, but may be more convenient to express. Predictor look-aheads reflect perfect information about the future, and may or may not be appropriate to the forecasting objectives. An algorithm will typically iterate through all valid origins within the training data to arrive at its preferred forecasting engine. Valid origins position themselves temporally in the data to leave enough distance to find the predictor data points prescribed by the farthest look-back and the reference data point prescribed by its look-ahead.

Statistical algorithms for constructing forecasting engines can accept predictor series directly in time series form, but machine learning algorithms operate on cross-sectional data with no notion of temporal spacing between data points. Hence, the data ensemble must be converted to cross-sectional form, explicitly coding temporal spacing between data points as extra features (dimensions) of the data points at all time steps at look-ahead distances and look-back distances from the origin. This implies converting to cross-sectional form anew for each origin used to train a forecasting engine.

An algorithm is instantiated by an algorithm class, set of hyper-parameters, and tuning rule. It outputs a forecasting engine that takes as input the same predictor series used in training, and outputs a forecast for a single time step. For forecasts that are longer than one time step, multiple forecasting engines may be necessary, each assuming a different reference look-ahead.

Forecast Construction Phase

In the forecast construction phase, a forecast covering a period is constructed by applying a forecasting engine or engines to predictor series within the testing data one time step at a time. The forecast assumes some valid origin. Any valid origin could be used, but using the earliest one leaves the most room for testing forecasts covering long periods.

There are several approaches available to construct a forecast covering some period from forecasts of single data points. We call fixing the origin and using several different forecasting engines, each assuming a different reference look-ahead, a direct extension rule. We call advancing the origin and using a single forecasting engine, which accepts as input data points previously forecasted, a recursive extension rule. With either extension rule, the origin can be advanced periodically to the time step just beyond the last forecasted time step – we call this the update cycle. At each origin advancement, or some integer multiple of origin advancements, the forecasting engines could potentially be re-constructed based on fresh, newly available training data – we call this the retrain cycle.

See more about extension rules in Appendix B.

To evaluate forecasting performance, forecasts are constructed over as many update cycles as are available in the testing data, and a metric is applied to the error of these forecasts with respect to the reference series. Note that evaluation is treated as part of the forecasting process, not a separate subsequent process, because how a forecast will be evaluated partially determines how it is constructed.

2.6.2 Decision Descriptions

We find it useful to categorize technique decisions into objective, form, and training and testing subsets.

Objective Decisions

Objective decisions serve the interests of the forecasting sponsor, but are not constrained to other particulars of the forecasting process. They specify what phenomenon is to be forecasted and how far into the future the forecast is to be made. In industry parlance, objective decisions are functional requirements, as distinguished from the engineering requirements. The job of the forecasting practitioner is then to accept objective decisions from the forecasting sponsor and make decisions about the rest of the forecasting process to produce appropriate forecasts.

Reference look-ahead	A period indicating how far from the origin a forecast should start.	a period
Update cycle	A period indicating how far the forecast should continue before the origin is advanced.	a period
Time step size	A period corresponding to the time step size used in evaluating forecasts. It can be different than the reference series or predictor series time step sizes, such as when making hourly forecasts based on data collected semi-hourly.	a period
Other objective decisions …		

Form Decisions

Form decisions determine how to construct the algorithm that will be used in turn to construct forecasting engines. They are typically made by the forecasting practitioner.

Span	Start and stop time stamps and the corresponding period indicating which contiguous subset of the reference and predictor series to use.	a datetime (start) to a datetime (stop)

Algorithm class	A functional form indicating how an algorithm uses hyper-parameters and cross-sectional data to construct forecasting models. Think of it as the learning engine within the algorithm.	A functional form
Hyper-parameters	Settings for an algorithm's hyper-parameters.	depends on algorithm class
Tuning rule	An indication of whether hyper-parameter setting selection is performed initially (static tuning) or deferred to forecast construction time (dynamic tuning). In dynamic tuning, a hyper-parameter selection criterion must be inferable.	`static tuning` or `dynamic tuning`
Reference data representation	An indication of how the reference series is represented, useful for analyzing the effects of different representations, e.g., integer coding schemes versus dummy variable schemes.	depends on reference data source
Predictor data representations	An indication of how predictor series are represented, useful for analyzing the effects of different representations, e.g., integer coding schemes versus dummy variable schemes.	depends on predictor data sources
Predictor look-backs	A set of vectors of periods indicating how far back from the origin to use predictor series data points. Each predictor series has its own vector of look-backs, which need not be the same length as those of other predictor series. A zero-value look-back refers to the origin. For example, a forecasting model that takes as inputs the value of electricity usage “E” at present and seven days ago, and the value of temperature “T” one day ago means the look-backs would be indicated by the set (E= (0 days, 7 days), T= (1 day)).	a set of vectors of periods, one vector for each predictor series

Predictor look-aheads	A set of vectors of periods indicating how far ahead of the origin to use predictor series data points. Using these data points as inputs implies that they are treated as forecasts themselves. Look-aheads are treated as decisions separate from look-backs for convenience; they could equivalently be expressed as negative look-backs.	a set of vectors of periods, one vector for each predictor series
Extension rule	An indication of whether a direct or recursive extension rule is used.	`direct` or `recurse`
Retrain cycle	A period indicating how often forecasting models should be re-trained, useful for analyzing the effect of training data freshness.	a period
Other form decisions …		

Training and Testing Decisions

Training and testing decisions are either prescribed by the forecasting sponsor or made by the forecasting practitioner.

In-sample/out-of-sample testing	An indication of whether testing is performed on the training or testing portion of the reference series. This allows for techniques that are properly tested on testing data to be compared to corresponding techniques improperly biased by training data to analyze the effects of potential overfitting.	`in sample` or `out of sample`
Holdout rule	An indication of how a data ensemble is partitioned into training and testing data. For example, a technique that reserves for testing the most recent 25% of data collected Jan 1 to Dec 31	`holdout specific` or `holdout last` or `cross-validation` or other rule

	would be indicated by a "holdout last" rule, with boundaries Oct 1 to Dec 31, and portion 25%.	
Holdout boundaries	Start and stop time stamps and the corresponding period indicating which contiguous subset of a data ensemble to not use as training data.	a datetime (start) to a datetime (stop)
Holdout	A fraction of a data ensemble not used as training data.	a fraction x in $0 < x < 1$
Clip	A fraction of the holdout not used as testing data. This allows for additional control over which data to use for testing. It is useful for analyzing the effect of temporal distance from the origin and to avoid anomalous data that may distort results. For example, a technique that ignores the most recent 5% of data collected to expose the influence of that data, perhaps because that data reflects a one-time event not relevant to forecasting, would be indicated by a 5% clip.	a fraction x in $0 \leq x < 1$
Other training and testing decisions …		

2.7 Metric Decisions

Metric decisions determine the operator used to quantify forecasting performance.

The part of the model of the forecasting process that addresses metric decisions is straight-forward: an operator is applied to a forecast and corresponding reference series to produce a measure of error between the two. Per the three-tier decision taxonomy, it assumes that data source, data strategy, and technique have been provided and are handled at higher levels, and

contents itself as a sub-process nested within the strategic and basic forecasting process models.

We treat penalty functions as custom metrics that value differences between actual and forecasted series per their economic impact. They can be defined to value such error asymmetrically, giving more or less weight to over-forecasted or under-forecasted levels. Further, they can take into account additional qualities, e.g., an implied reserve amount that limits economic damage caused by under-forecasting.

Metrics are chosen to correspond to some characteristic of forecasts important to the forecasting sponsor. They are prescribed by the forecasting sponsor or made by the forecasting practitioner.

Metric (or penalty function)	An operator on two time series used to score their difference. One time series represents a portion of the actual reference series, the other represents the corresponding predicted portion of the reference series. The time series need not be the same length, provided they are temporally aligned and the operator can sensibly compare levels in one time series with null values in the other.	an operator

2.8 Methods

With our models of the forecasting process in hand, we introduce methods to analyze forecasting process decision effects on forecasting performance.

2.8.1 General Method to Analyze Decisions

Our method to analyze forecasting process decision effects, in its most general form, comprises five steps.

1. Explicate decisions and decision options.
2. Exhaustively vectorize all decision option combinations.
3. Use the vectors to instantiate specific forecasting processes and construct forecasts.
4. Apply metrics to score the forecasts.
5. Map the scores back to the decisions to analyze their effects.

2.8.2 Method Variation to Analyze Technique Decisions

Our method variation to analyze technique decisions is a special case of the general method, and is based on the basic forecasting process model.

1. **Scope the analysis.**
    a. Choose a set of forecasting process decisions that could potentially affect forecasting performance and are of interest to study.
    b. Choose a set of forecasting techniques to study by systematically instantiating the model over a range of options for each decision, such that each technique is uniquely specified by a vector of length equal to the number of decisions, and each vector element is equal to one decision option. Restrict the range of options for each prescribed decision appropriately. Choose a range of options for each flexible decision.
    c. Choose a set of metrics to study.
2. **Prepare the data.**
    a. Gather data sources for the reference series and predictor series as indicated by the relevant decisions.

b. Pre-process the data sources. Transform the data to representations indicated by the relevant decisions, if necessary. Transform the data to time series form, if necessary.

c. Aggregate the data. Consolidate time series constituting the reference series by sum, mean, or other aggregator, as appropriate. Similarly, consolidate time series constituting predictor series.

d. Compile data ensembles, one ensemble for each set of techniques sharing span and time step size options. Each data ensemble then comprises data describing the reference series and predictor series in time series form, aligned to the same start date, stop date, and time step size.

3. **Examine the data.** Note data characteristics of individual series and inter-series relationships.

4. **Construct forecasts and evaluations.** Systematically invoke techniques and metrics to construct forecasts and evaluate forecasting performance.

5. **Examine the results.**

    a. Examine inter-metric relationships. Specifically, examine how the distribution of scores as measured by one metric correlates with those of other metrics. Choose metrics on which to focus the remainder of the analysis.

    b. Examine metric score distributions and rank distributions – across all techniques. Rank techniques per their scores per one or more metrics using some ranking rule. Note relationships within and among the distributions.

    c. Examine metric score distributions and rank distributions – across qualified techniques. Choose qualification thresholds for one or more metrics, such that any technique is qualified only if its scores cross the thresholds for each of the metrics. If no techniques are qualified per these thresholds, reconsider the choice of metrics and their qualification thresholds, accordingly. Rank qualified techniques and note relationships within and among the distributions.

2.8.3 Method Variation to Analyze Data Strategy Decisions

The method variation to analyze technique decisions can be applied to data strategy decisions. For deeper analysis of data strategy decisions, though, we formulate another variation that exposes trends in forecasting performance across sample sizes, numbers of clusters, and time step sizes, for data strategies that adopt any one of the basic integration rules.

1. **Scope the analysis.**
    a. Choose a set of forecasting process decisions, including data strategy decisions, to study.
    b. Choose a set of forecasting techniques to study by systematically instantiating the model over a range of options for each data strategy decision. Restrict the range of options for each prescribed decision and non-data strategy decision appropriately. Choose a range of options for each data strategy decision.
    c. Choose a set of metrics to study.
2. **Prepare the data.**
    a. Gather data sources.
    b. Pre-process the data sources.
    c. Organize data into samples, if indicated by data strategy decisions to do so.
    d. Organize data into groups of clusters, if indicated by data strategy decisions to do so.
    e. Organize data into temporally magnified representations, if indicated by data strategy decisions to do so.
    f. Aggregate the data. Consolidate time series constituting each sample or cluster corresponding to the reference series. Similarly, consolidate time series constituting each sample or cluster corresponding to the predictor series.
    g. Compile data ensembles, one ensemble for each sample or cluster. Each data ensemble then comprises data describing the reference series and predictor series in time series form for a specific sample or cluster.
3. **Examine the data.**

4. **Construct forecasts and evaluations.** Systematically invoke data strategies, techniques, and metrics to construct forecasts and evaluate forecasting performance, taking care to include any special treatment indicated by the integration rule decision.
5. **Examine the results.** Examine metric score distribution trends and rank distribution trends across sample sizes, numbers of clusters, and time step sizes – across all techniques.

2.9 Computation Platform

To practically realize our analysis methods, we develop an original computation platform. The current version is implemented in R and comprises about 5,000 lines of code. [72] It leverages the *rminer* and other popular machine learning packages, leverages the *relaimpo* package for relative importance scoring, and makes extensive use of the *ggplot2* package for custom data visualizations. [68,69,70,72,73] Computation platform functionality includes pre-processing specific electricity usage data; pre-processing temperature data; generating calendar data; grouping data by sampling, clustering, and temporal magnification strategies; aligning data from different sources into data ensembles; managing data formats; computing forecasts and metric scores; and visualizing results in a variety of standard and custom graphics.

2.9.1 Design

The design of the computation platform is guided by the three-tier decision taxonomy and forecasting process models. As such, an evaluation machine (that handles metric decisions) is nested within a forecasting/evaluation machine (that handles technique decisions), which in turn is nested within a strategic forecasting/evaluation machine (that handles data source and data strategy decisions).

The evaluation machine is straightforward. The machine accepts as input a forecast, some temporally aligned test data, and a metric that operates on the forecast and test data. It outputs a score quantifying the forecast performance (Figure 2-5).

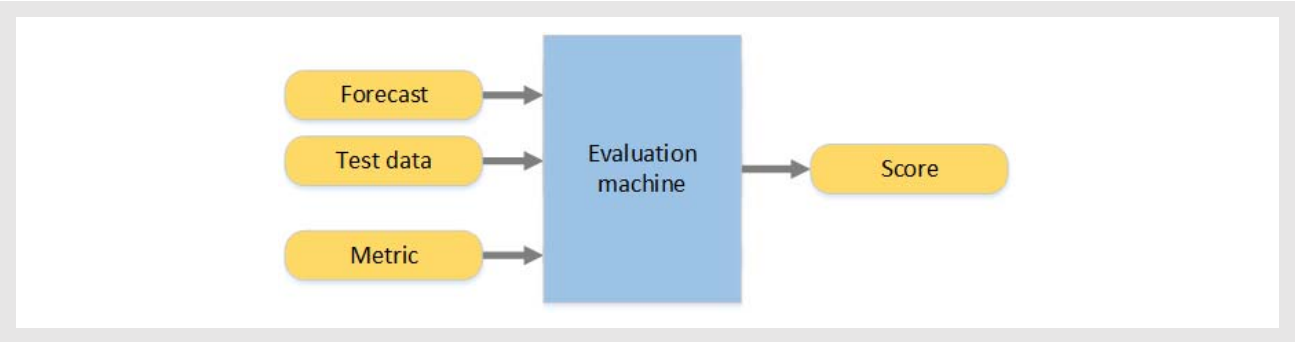


Figure 2-5: Evaluation machine. *Blue* indicates the machine. *Yellow* indicates inputs and outputs.

The forecasting/evaluation machine incorporates the evaluation machine just described, and a forecasting machine. The forecasting machine accepts as input a technique, which instantiates a specific forecasting process to be realized by the machine. The technique includes pointers to data sources, so the forecasting machine is provided everything it needs to unambiguously construct and output a forecast and corresponding test data (Figure 2-6).

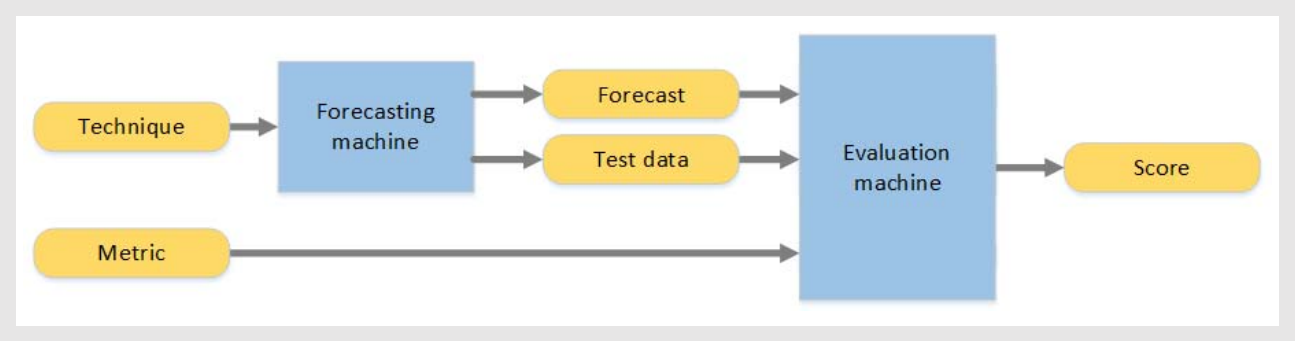


Figure 2-6: Forecasting/evaluation machine. *Blue* indicates component machines. *Yellow* indicates inputs and outputs.

The strategic forecasting/evaluation machine incorporates both the evaluation and forecasting machines just described, and additionally data manipulation and forecast consolidation machines. The data manipulation machine accepts as input a data strategy and the part of a technique that points to data sources, and outputs the data sources appropriately transformed

to a sample, group of clusters, temporal magnification, or other form. Each of these outputs is in turn provided as input to the forecasting machine, along with a technique, which outputs a set of forecast/test data pairs, one pair corresponding to each transformed data source input. At this step, the test data are discarded. All the forecasts as a group are provided as input to the forecast consolidation machine, which outputs a single forecast/test data pair. This forecast and test data, plus a metric, are provided as input to the evaluation machine, which finally outputs a score (Figure 2-7).

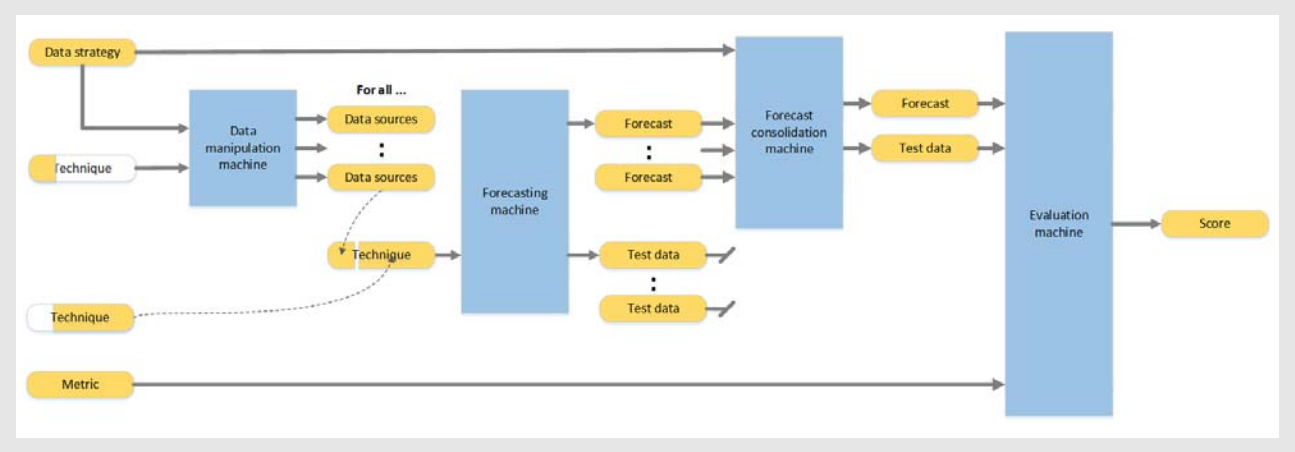


Figure 2-7: Strategic forecasting/evaluation machine. *Blue* indicates component machines. *Yellow* indicates inputs and outputs.

2.9.2 Architecture and Data Organization

The computation platform is architected as five modules, which communicate with each other in a variety of data formats.

Code Modules	• Module 1 – Pre-process • Module 2 – Group • Module 3 – Ensemble • Module 4 – Experiment • Module 5 – Visualize

Data File Formats	• Raw CER format • Raw SGSC format • multiID format • uniID format • Ensemble format • Technique format • Result format (forecast, technique, evaluation, compute time) • Result format (fast evaluation)

2.9.3 Computing Functionality

The computation platform includes much original and leveraged code to provide extensive functionality.

Manage data formats	• Convert between data formats
Manage data scale	• Auto-parallelize data operations
Pre-process data	• Pre-process Ireland CER data • Pre-process Australia SGSC data • Pre-process WU temperature data • Pre-process airport temperature data • Generate calendar data • Find clean data range • Handle daylight savings time • Impute missing data
Group data by samples	• Re-sample by bootstrap version 1 • Re-sample by bootstrap version 2 • Re-sample by bootstrap version 3 • Re-sample by jackknife
Group data by clusters	• Cluster by geographic proximity • Cluster by electricity usage correlation • Cluster by electricity usage coherence
Build data ensemble	• Coarsen/refine source data resolution • Align source data

Compute forecasts	• Convert time-series data to cross-sectional data • Extend forecast by extension rule • Apply machine learning algorithms
Score forecasts	• Score forecast
Rank forecasts	• Rank by single metric • Rank by multiple metrics

2.9.4 Supported Algorithms

The computation platform supports custom algorithms (algorithm class + its hyper-parameters), provided as any R functions parameterized to accept an aligned data ensemble representing the reference series and predictor series. For convenience, the current version provides the following already-implemented standard algorithms.

knn	Nearest neighbor	k=5, kernel=rectangular
linreg	Linear regression	
mlp	Multilayer perceptron	1 hidden layer, normalized, weight initialization=0.7, decay=0, maximum iterations=100, absolute tolerance=0.0001, relative tolerance=$1x10^{-8}$, activation function=sigmoid
naïve	Naïve	forecasted level is always same as origin level
svm	Support vector regression	normalized, scaled, type=eps, kernel=radial, gamma=1.0, cost=1000, tolerance=0.01, epsilon=0.1
tree	Decision tree	method=anova, complexity parameter=$1x10^{-10}$

2.9.5 Supported Metrics

The computation platform supports custom metrics, provided as any R functions parameterized to accept two same-size numeric vectors representing aligned forecast and reference series. For convenience, the current version provides the following already-implemented metrics.

SAE ↓	sum absolute error/deviation	[0, ∞]
MAE ↓	mean absolute error	[0, ∞]
MdAE ↓	median absolute error	[0, ∞]
GMAE ↓	geometric mean absolute error	[0, ∞]
MaxAE ↓	maximum absolute error	[0, ∞]
RAE ↓	relative absolute error	[0%, ∞]
SSE ↓	sum squared error	[0, ∞]
MSE ↓	mean squared error	[0, ∞]
MdSE ↓	median squared error	[0, ∞]
RMSE ↓	root mean squared error	[0, ∞]
GMSE ↓	geometric mean squared error	[0, ∞]
HRMSE ↓	heteroscedasticity consistent root mean squared error	[0, ∞]
RSE ↓	relative squared error	[0, ∞]
ME ↓	mean error	[0, ∞]
SMinkowski3 ↓	sum of Minkowski loss function (q=3, heavier penalty for large errors when compared with SSE)	[0%, ∞]
MMinkowski3 ↓	mean of Minkowski loss function (q=3, heavier penalty for large errors when compared with SSE)	[0%, ∞]
MdMinkowski3 ↓	median of Minkowski loss function (q=3, heavier penalty for large errors when compared with SSE)	[0%, ∞]
COR ↑	correlation	[-1, 1]
q2 ↓	1-correlation^2 test error metric, as used by M.J. Embrechts	[0, 1]

R2 ↑	coefficient of determination R^2 (squared Pearson correlation coefficient)	[0, 1]
Q2 ↓	R^2/SD test error metric, as used by M.J. Embrechts	[0, ∞]
NAREC ↑	normalized REC area (given a fixed val=tolerance)	[0, 1]
TOLERANCE ↑	the tolerance (y-axis value) of a REC curve (given a fixed val=tolerance)	[0, 1]
MAPE ↓	mean absolute percentage error	[0%, ∞]
MdAPE ↓	median absolute percentage error	[0%, ∞]
RMSP ↓	root mean square percentage error	[0%, ∞]
RMdSPE ↓	root median square percentage error	[0%, ∞]
SMAPE ↓	symmetric mean absolute percentage error	[0%, 200%]
NRMSE ↓	normalized root mean square error	[0, 1]
MASE ↓	Mean absolute scaled error	[0, ∞]

Additionally, also for convenience, the current version provides 30 already-implemented penalty functions that can be used as metrics.

OFT.r0 ↓ OFT.r10 ↓ OFT.r25 ↓ UFT.r0 ↓ UFT.r10 ↓ UFT.r25 ↓	fraction of time that the technique over-forecasted or under-forecasted, assuming a reserve of 0%, 10%, or 25%	[0%, 100%]
OFM.r0 ↓ OFM.r10 ↓ OFM.r25 ↓ UFM.r0 ↓ UFM.r10 ↓ UFM.r25 ↓	fraction of the amount that the technique over-forecasted or under-forecasted, assuming a reserve of 0%, 10% or 25%.	[0%, ∞]

PT.r0.1.1 ↓ PT.r0.2.1 ↓ PT.r0.1.2 ↓ PT.r10.1.1 ↓ PT.r10.1.2 ↓ PT.r10.2.1 ↓ PT.r25.1.1 ↓ PT.r25.1.2 ↓ PT.r25.2.1 ↓	weighted average of the over-forecasts and under-forecasts, with respect to time, assuming a reserve of 0%, 10%, or 25%, and weighted 1 or 2 for over-forecasts and 1 or 2 for under-forecasts	[0%, 100%]
PM.r0.1.1 ↓ PM.r0.1.2 ↓ PM.r0.2.1 ↓ PM.r10.1.1 ↓ PM.r10.1.2 ↓ PM.r10.2.1 ↓ PM.r25.1.1 ↓ PM.r25.1.2 ↓ PM.r25.2.1 ↓	The weighted average of the over-forecasts and under-forecasts, with respect to amount, assuming a reserve of 0%, 10% or 25%, and weighted 1 or 2 for over-forecasts and 1 or 2 for under-forecasts	[0%, ∞]

2.9.6 Data Visualization Functionality

The computation platform provides a variety of custom data visualizations.

Series visualizations (full population, sample, or cluster)	• Visualize series see Figure 4-3 • Visualize series trends • Visualize series seasonality • Visualize series correlations see Figure 4-4 • Visualize series autocorrelation see Figure 4-5 • Series geo-location map see Figure 6-2
Forecast visualizations (full population, sample, or cluster)	• Visualize forecast • Visualize forecast error • Visualize forecasts as overlay see Figure 4-6

	• Visualize forecast error as overlay see Figure 4-7
Technique score visualizations	• Visualize technique score distribution see Figure 4-11 • Visualize relative importance of decisions to technique score see Figure 4-8 • Visualize technique scores with decision option trend see Figure 4-16 • Visualize technique scores with algorithm competition see Figure 4-12 • Visualize monotonicity • Visualize best technique score distribution by data source see Figure 3-4 • Visualize best technique algorithm class-family distribution by data source see Figure 3-5, Figure 3-10
Metric relationship visualizations	• Visualize metric mean score correlations see Figure 4-18 • Visualize techniques by rank across metrics, scatter • Visualize techniques by rank across metrics, stack
Technique rank visualizations	• Visualize techniques by rank see Figure 4-20, Figure 4-21, Table 4-1 • Visualize techniques by decision option distribution see Figure 4-22 • Visualize technique families by rank split by decision options see Figure 4-26 • Visualize technique families by rank split by decision options in context see Figure 4-35 • Visualize techniques by rank with multiple metrics see Table 6-7
Sample visualizations	• Visualize sample score trend across sample sizes see Figure 5-8

	• Visualize sample rank trend across sample sizes see Figure 5-10 • Visualize sample score impact trend, multiple locations see Figure 6-53 • Visualize bounded sample score impact trend, multiple locations see Figure 6-54
Cluster visualizations	• Visualize series mean correlations • Visualize cluster assignment, phylogram see Figure 5-5 • Visualize cluster score trend across cluster counts see Figure 5-20 • Visualize cluster rank trend across cluster counts see Figure 5-22 • Visualize cluster count score impact trend, multiple locations see Figure 6-57 • Visualize bounded cluster count score impact trend, multiple locations see Figure 6-58
Temporal magnification visualizations	• Visualize temporal magnification score trend across time step sizes see Figure 5-27 • Visualize temporal magnification rank trend across time step sizes see Figure 5-29 • Visualize temporal magnification score impact trend, multiple locations see Figure 6-60 • Visualize bounded temporal magnification score impact trend, multiple locations see Figure 6-61

2.10 Proof-of-Concept

To demonstrate our analysis methods for metric and technique decisions, we apply them here to explore relationships between data source characteristics and forecastability, using the 3,003 benchmark time series from the M3 Forecasting Competition as reference and predictor series. [156]

(Further demonstration of methods to analyze technique and data strategy decisions follow in subsequent chapters.)

2.10.1 Metric Decisions

To analyze the effects of a metric decision ranging over 3 options, we construct forecasts for all 3,003 series using the same one technique, and evaluate each per the 3 metrics.

Reference data source	3,003 options	From M3 Forecasting Competition. Any of 3,003 anonymous time series datasets of length 20 to 144 time steps.
Predictor data source	3,003 options	*Same as reference data source.*
Algorithm class	lock-in	support vector regression
Extension rule	lock-in	direct
Update cycle	lock-in	1 time step
Holdout	lock-in	0.50
Metric	3 options	MAPE, MASE, R2

We then rank order the data sources by their associated forecast scores per one metric, MAPE, color coding data sources based on their spans, to show us the relative forecastability of the data sources (Figure 2-8). Those near the front, because their forecasts evaluate to low MAPE

scores, lend themselves to being forecasted by the technique. We see that most of the long data sources gravitate to the front, and so we may suspect a relationship between data span and forecastability.

However, when we compare data sources evaluated per the other two metrics, MASE and R^2, but still rank ordered per MAPE score, we see very different patterns. Many of the data sources highly forecastable in the context of a MAPE metric decision are poorly forecastable in the context of a MASE or R^2 metric decision. MAPE correlates to MASE at $r = 0.15$, MASE to R^2 at $r = -0.20$, and MAPE to R^2 at $r = -0.09$. Across all metrics, 85.23% of metric-pairs show statistically significant correlation, with p-value < 0.05, albeit most of these correlations are low (Figure 2-9).

Forecastability is highly sensitive to the metric decision.

2.10.2 Technique Decisions

To analyze the effects of technique decisions, we vary four technique decisions, apply the resultant techniques to construct 32 forecasts for each of the 3,003 series, and evaluate each per one metric, MAPE.

Reference data source	3,003 options	From M3 Forecasting Competition. Any of 3,003 anonymous time series datasets of length 20 to 144 time steps.
Predictor data source	3,003 options	*Same as reference data source.*
Algorithm class	4 options	linear regression, multilayer perceptron, naïve, support vector regression
Extension rule	2 options	direct, recurse
Update cycle	2 options	1 time step, 2 time steps

Holdout	2 options	33%, 50%
Metric	lock-in	MAPE

We see that each series is associated with a distribution of MAPE scores, some ranging from near zero to over 100. Correlations between techniques range from r = 0.30 to r = 1.0. Most technique-pairs are correlated strongly, and all correlations are statistically significant, with p-value < 0.05 (Figure 2-10).

Looking at the mean or best scores for each series, or looking at the scores for one specific technique for each series, give different pictures of forecasting performance (Figure 2-11).

We can explore the relative performance of various algorithm classes, too. In the context of locked-in decisions for all but algorithm class, we identify the best performing techniques by algorithm class for each series, and see a distinctive distribution of the winning algorithm classes across the series – naïve wins in 59% of series, linear regression wins in 38% of series (Figure 2-12). In the context of many decisions, we see a somewhat different distribution of winning algorithm classes – naïve wins in only 50% of series, linear regression wins in almost 41% of series. (Figure 2-13). In this case, which algorithm class ranks best does not change in the context of many decisions, but we do see that the best performing algorithm class is not quite so much better as we would otherwise have thought. Further, forecasts constructed by the winning algorithm classes score 15.26 MAPE points better on average when considered in the context of many decisions. So, we get a richer picture of just how good the good algorithm classes are.

Locked-in decisions can distort the view of forecasting performance.

Locked-in decisions can distort the view of algorithm class importance to forecasting performance.

2.11 Tables and Data Visualizations

Table 2-1:
Electricity demand forecasting research classified by look-ahead, based on Hong. [18]

Classification		Predictor Variables			Lookahead	Update Cycle
		Temperature	Economics	Land Use		
VSTLF	Very Short-Term Load Forecasting	optional	optional	optional	≤ 1 day	≤ 1 hour
STLF	Short-Term Load Forecasting	required	optional	optional	1 day to 2 week	1 hour to 1 day
MTLF	Medium-Term Load Forecasting	simulated	required	optional	2 week to 3 year	1 day to 1 month
LTLF	Long-term Load Forecasting	simulated	simulated	required	3 year to 30 year	1 month to 1 year

Table 2-2:
Electricity demand forecast uses, based on Hong. [18]

Utility Use	VSTLF	STLF	MTLF	LTLF
Energy purchasing	☑	☑	☑	☑
Transmission & distribution planning		☑	☑	☑
Operations	☑	☑		
Demand-side management	☑	☑	☑	☑
Financial planning			☑	☑

Table 2-3: Popular algorithm classes (statistical linear).

CLASS	Sub-Class	Algorithm	
		Univariate	Multivariate
Statistical Linear	Baseline	Naïve Mean Seasonal Naïve	
	Linear Regression	Linear Regression	Multivariate Linear Regression
	Exponential Smoothing	Linear Trend Simple Exponential Smoothing (SES) Holt Linear Exponential Trend Additive Damped Trend Multiplicative Damped Trend Holt-Winters Additive Holt-Winters Multiplicative Holt-Winters Damped Various other exponential smoothing Exponential Smoothing with Discontinuities Exponential Smoothing with Constraints Exponential Smoothing with Renormalization of Seasonal Indices Simple Exponential Smoothing with Drift3 – Theta Croston's	Multivariate Simple Exponential Smoothing
	ARIMA-type	Autoregression (AR) Moving Average (MA) ARMA ARIMA Seasonal ARIMA ARARMA AR with Moving Horizon Dynamic Regression Integer-Valued Moving Average	Vector Autoregression (VAR) Bayesian Vector Autoregression (BVAR) ECM Asymmetric Linear Vector ECM Vector ARMA (VARMA) Vector ARIMA (VARIMA)
	ARIMA-type/Long Memeory ARIMA-type/Long Memeory ARIMA-type/Long Memeory	Fractionally Differenced ARMA (FARMA) Fractionally Integrated ARMA (ARFIMA) Periodic ARFIMA	Multivariate Fractionally Integrated ARMA (Multivariate ARFIMA)
	State Space	Balanced State Space Single Source of Error State Space Other ETS per state space models	
	State Space/Dynamic Linear (DLM)	Basic Structural (BSM) Linear Growth Structural Continuous-time Structural Non-Gaussian Structural: Power Steady	

Table 2-4: Popular algorithm classes (statistical non-linear and machine learning).

CLASS	Sub-Class	Algorithm	
		Univariate	Multivariate
Statistical Non-Linear	Bilinear	Bilinear	
	Regime Switching	Regression Splines	Threshold Vector Autoregression
		Piecewise Linear Regression	
		Threshold Autoregression (TAR)	
		Self-Exciting Threshold Autoregression (SETAR)	
		Continuous-time Threshold Autoregression (CTAR)	
		Smooth Transition Autoregression (STAR)	
		Multi-level Panel Smooth Transition Autoregression	
		Markov Regime Switching	
		Functional Coefficient (FCAR)	Vector Functional Coefficient (VFCAR)
	ARCH-type	Autoregressive Conditional Heteroscedastic (ARCH)	
		Generalized Autoregressive Conditional Heteroscedastic (GARCH)	
		Fractionally Integrated GARCH (FIGARCH)	
	Quantile Regression	Conditional Quantile Regression	Multivariate Conditional Quantile Regression
	Non-Guassian Time Series	Autoregressive Conditional Duration	
	Other Statistical Non-Linear	Fourier Coefficients	
		Flexible Non-Linear Regression	
		Projection Pursuit Regression	
		Time-varying Autoregression	
		Random Walk with Drift	
		Hierarchical Time Series	
		Inhomogeneous Poisson Regression	
		Singular Value Decomposition	
		Conditional Restricted Boltzman Machine	
Machine Learning	Logistic Regression		Logistic Regression
			Regression with Regularization
	Support Vector		Support Vector Regression (SVR)
	Artificial Neural Network (ANN)		Multilayer Perceptron (MLP)
			Single Hidden Layer Feedforward Neural Network (SLFN)
			Extreme Learning Machine (ELM)
			Recurrent Extreme Learning Machine (RELM)
			Bayesian Neural Network
			Radial Basis Function (RBF)
			Generalized Regression Neural Network (GRNN)
			Adaptive Neural-based Fuzzy Inference System (ANFIS)
			Neural Network Autoregression
	Probablistic		Naïve Bayes
	Similarity		K-Nearest Neighbor Regression (kNN)
			Lazy Learning
	Decision Tree-based		Classification and Regression Trees (CART)
			Random Forest
			Ensemble of trees
	Rule-based		Rule Set
	Stochastic		Genetic Programming (GP)
			Gaussian Process Regression (GP)
			Stochastic Gradient Descent
	Enhancements		Gradient Boosting
			Adaptive Boosting

Table 2-5: Popular hyper-parameter selection criteria.

CLASS	↳	Hyper-Parameter Selection Criterion
Akaike	↳	Akaike's Information Criterion (AIC)
		Akaike's Final Prediction Error (FPE)
Bayes		Bayes Information Criterion (BIC)

Table 2-6: Popular data transformations.

CLASS	↳	Data Pre-Process Transformation
Time Series		Time Series Differencing Transformation
		Moving Average Transformation
		Log Transformation
		Deseasonalization
		X-11
		X-11-ARIMA
		X-12-ARIMA
		TRAMO-SEATS
		STL
		Sinusoidals
		Combined seasonal components
		Unit roots
		Scaling
Machine Learning	↳	Time Series to Cross-Sectional
	↳	Scaling
	↳	Normalization

Table 2-7: Popular extension rules.

CLASS	↳	Strategy
Direct	↳	Direct
		Direct with Updates
		Direct with Continuous Updates
Recursive		Recursive
		Recursive with Updates
Combination		Direct with Continuous Updates
		Direct + Recursive Combination (DirRec)
		Multi-Input Multi-Output (MIMO)
		Direct + MIMO Combination (DIRMO)
		Iterated PRESS5
		RECNOISY5

Table 2-8: Popular training and testing rules.

CLASS	↳	Data Apportionment	
Holdout		Holdout Last (general)	Train on first (1-x) of time series data; test on last x of time series data.
	↳	Holdout Last (1/3)	Train on first 2/3 of time series data; test on last 1/3 of time series data.
		Holdout Similar	Train on portion of time series data that is seasonally similar to test portion of time series data.
Cross-validiation		Cross-sectional Cross-validation	Train on arbitrary portion of cross-sectional data, repeat x times and take mean.
		Time Series Cross-validation	
Split Sample Validation		Split Sample Validation	

Table 2-9: Popular metrics.

CLASS	ϟ	Metric	
Correlation		r, COR	Correlation Coefficient
		R2	Coefficient of determination
		Adj R2	Adjusted R Squared
		q2	1-r2
		Q2	R2/standard deviation
Absolute Error		GMAE	geometric mean absolute error
		MdAE	Median Absolute Error
		MSE	MSE
		NAREC	normalized REC area at fixed tolerance
		SAE	sum absolute error/deviation
		SSE	sum squared error
Relative Error		GMRAE	Geometric Mean Relative Absolute Error
		GMSE	geometric mean squared error
		HRMSE	Heteroscedasticity consistent root mean squared error
		MAE	Mean Absolute Error
		MaxAE	maximum absolute error
	ϟ	MAPE	Mean Absolute Percentage Error
		MdAPE	Median Absolute Percentage Error
		MdSE	median squared error
		ME	mean error
		MRAE	Mean Relative Absolute Error
		MdRAE	Median Relative Absolute Error
		MdMinkowski3	median of Minkowski loss function
		MMinkowski3	mean of Minkowsky loss function
	ϟ	NRMSE	Normalized root mean square error
		RAE	relative absolute error
		RelMAE	Relative Mean Absolute Error
		RelRMSE	Relative Root Mean Squared Error
		RSE	relative squared error
	ϟ	RMSE	Root Mean Square Error
		RMdSPE	Root Median Square Percentage
		RMSPE	Root Mean Square Percentage
		RRSE	root relative squared error
		sMAPE	Symmetric Mean Absolute Percentage Error
		sMdAPE	Symmetric Median Absolute Percentage Error
		SMinkowski3	sum of Minkowski loss function
		Tolerance	tolerance of REC curve at fixed tolerance
Other		E	Nash-Sutcliffe Efficiency Coefficient
		GFESM	Generalized Forecast Error Second Moment
		LMR	Log Mean Squared Error Ratio
		MASE	Mean Absolute Scaled Error
		PB	Percentage Better
		PBMAE	Percentage Better Mean Absolute Error
		PBMSE	Percentage Better Mean Squared Error
		Residuals Diagnostics1	Residuals Diagnostics
		Residual Autocorrelation1	Residual Autocorrelation
Cost		Cost	Cost given an arbitrary cost function

Table 2-10: Integration rules at fixed time step size.

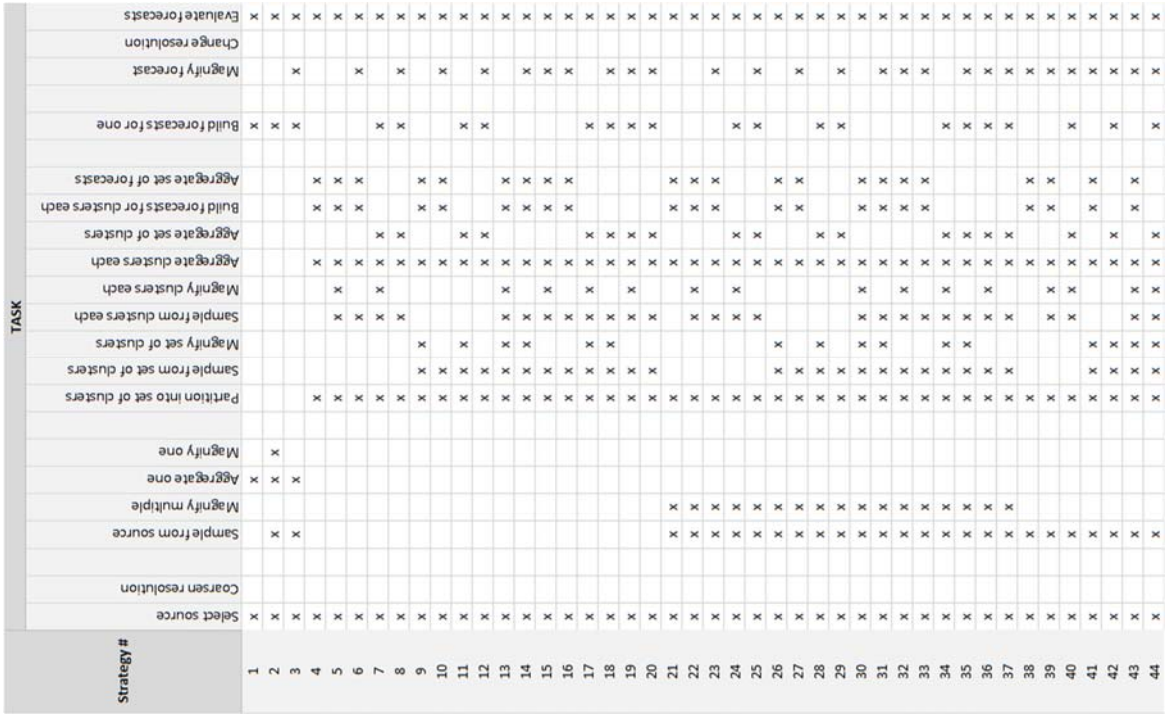


Table 2-11: Integration rules at refining time step size.

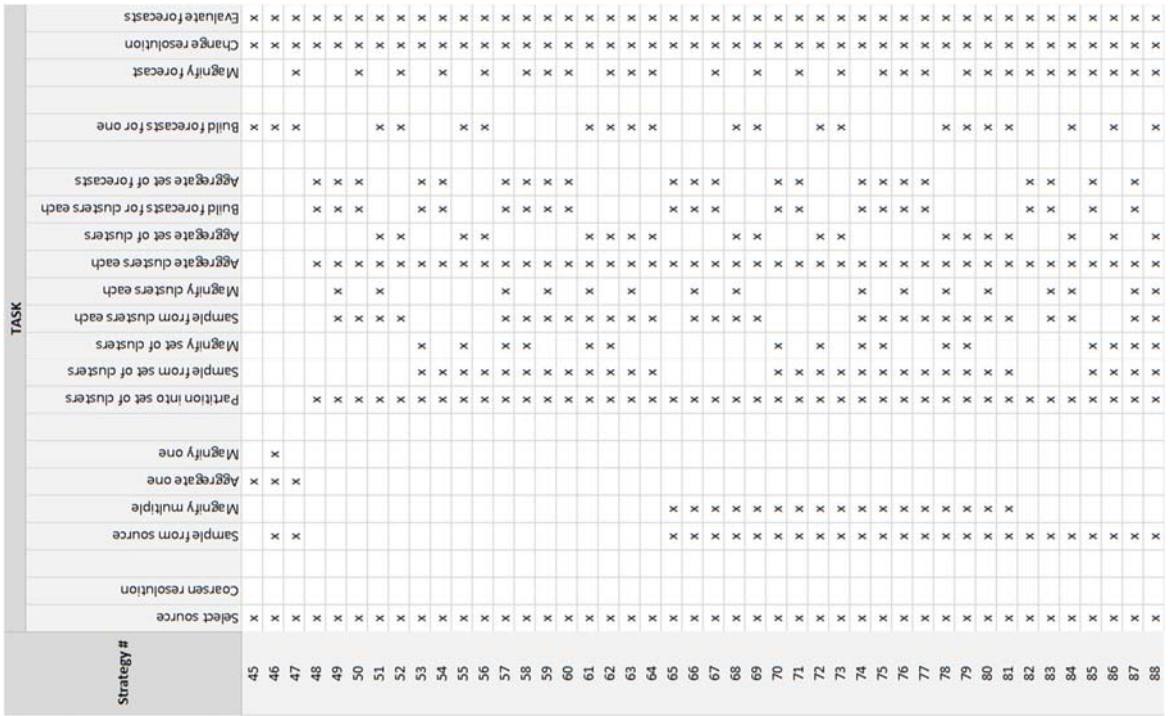
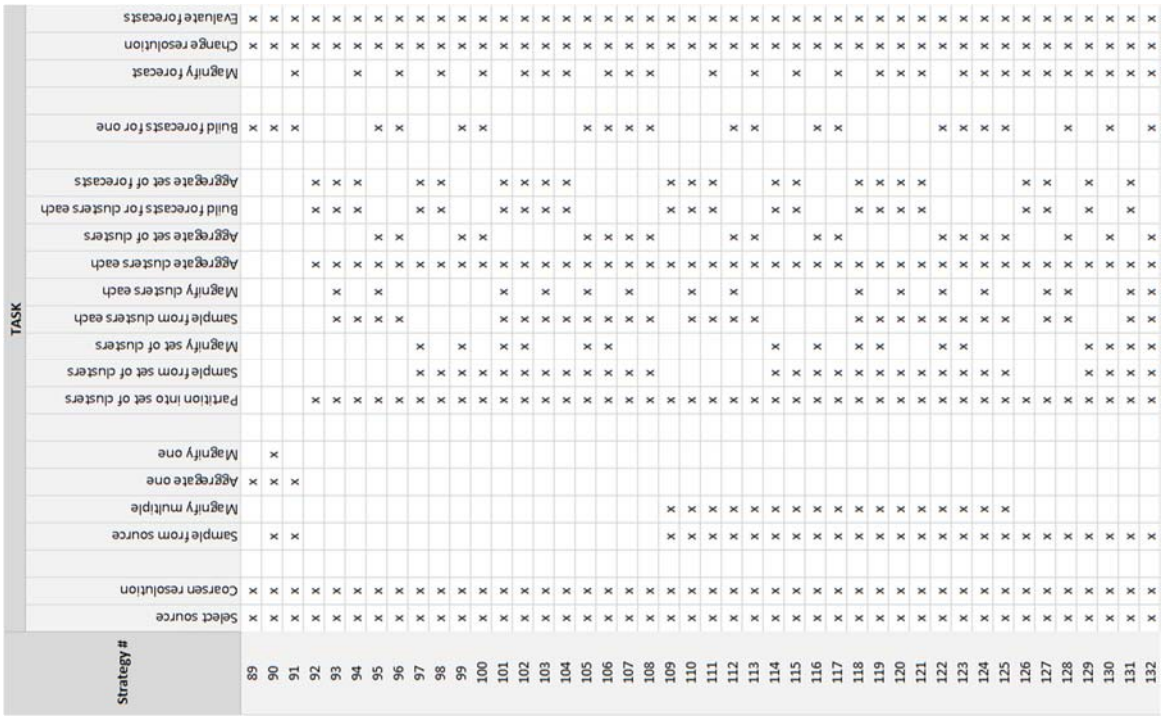


Table 2-12: Integration rules at coarsening time step size.



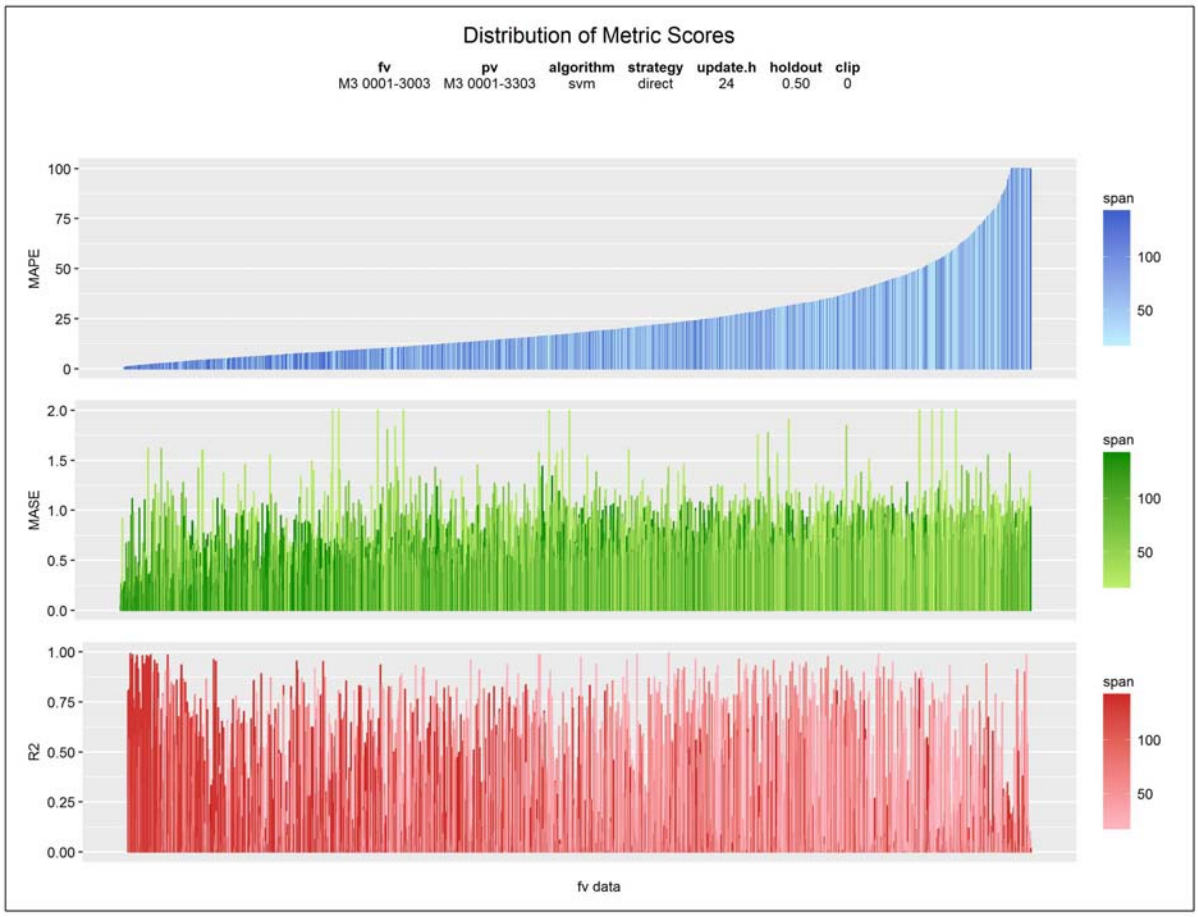


Figure 2-8: Distribution of metric scores across data sources. 3,003 data sources from the M3 Forecasting Competition. Data sources are arranged along x-axis by rank order per MAPE score of associated forecast made by a specific technique. *Blue* is MAPE score, *green* is MASE score, *red* is R^2 score. *Darker bars* are longer data spans, *lighter bars* are shorter data spans.

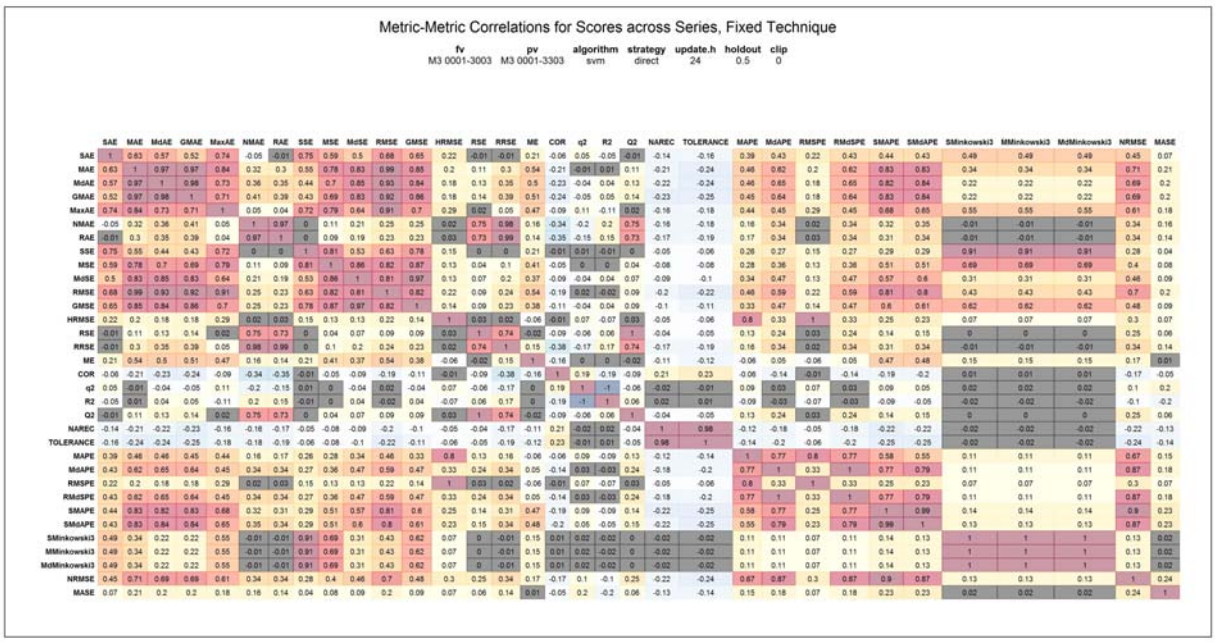


Figure 2-9: Metric-metric correlations for metric scores across data sources, fixed technique. 1 technique. 3,003 data sources from the M3 Forecasting Competition. For each cell, all data sources are scored per two metrics, and the sequence of scores per the first metric is correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation. *Gray* is statistically not significant, p-value < 0.05.

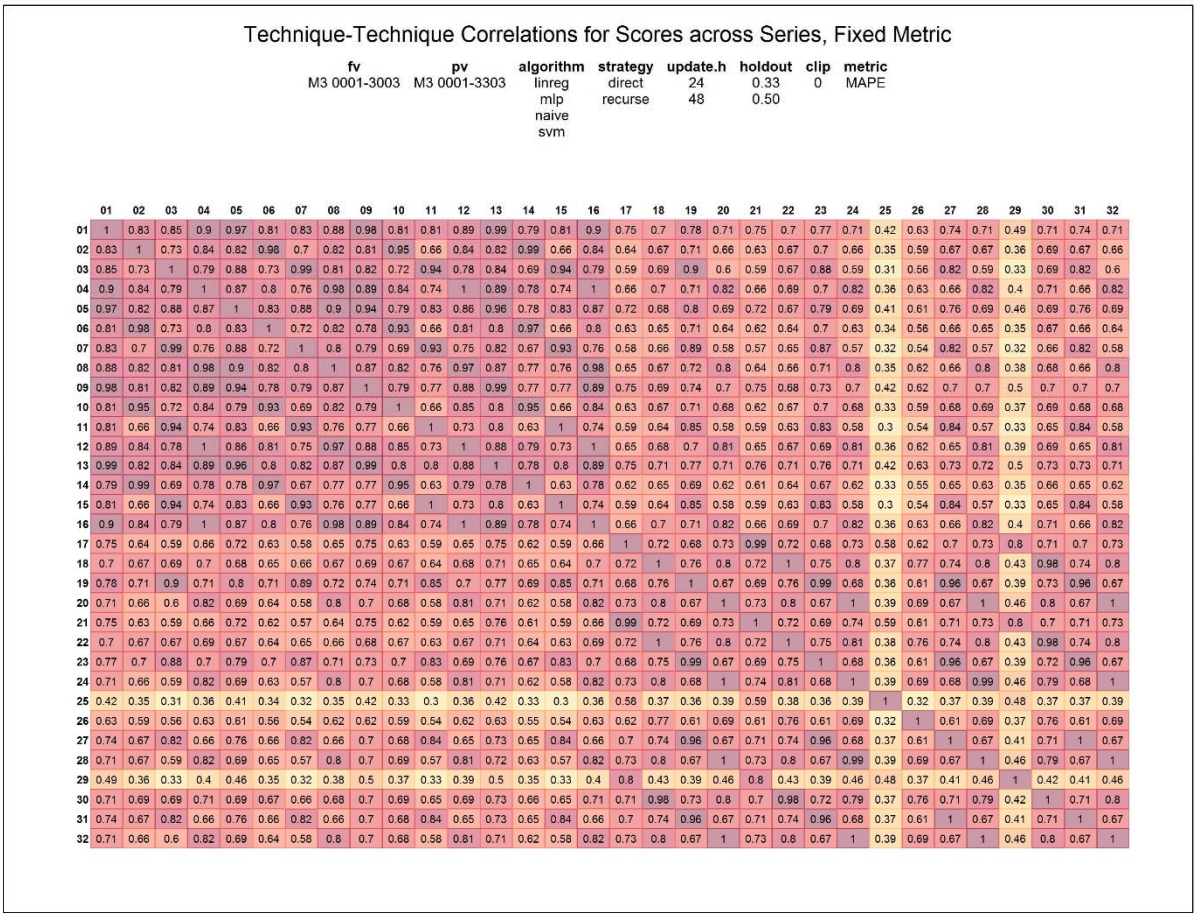


Figure 2-10: Technique-technique correlations for metric scores across data sources, fixed metric. 32 techniques. 3,003 data sources from the M3 Forecasting Competition. Metric is MAPE. For each cell, all data sources are scored per two techniques, and the sequence of scores per the first technique is correlated with the sequence of scores per the second technique. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation. All correlations are statistically significant, p-value < 0.05.

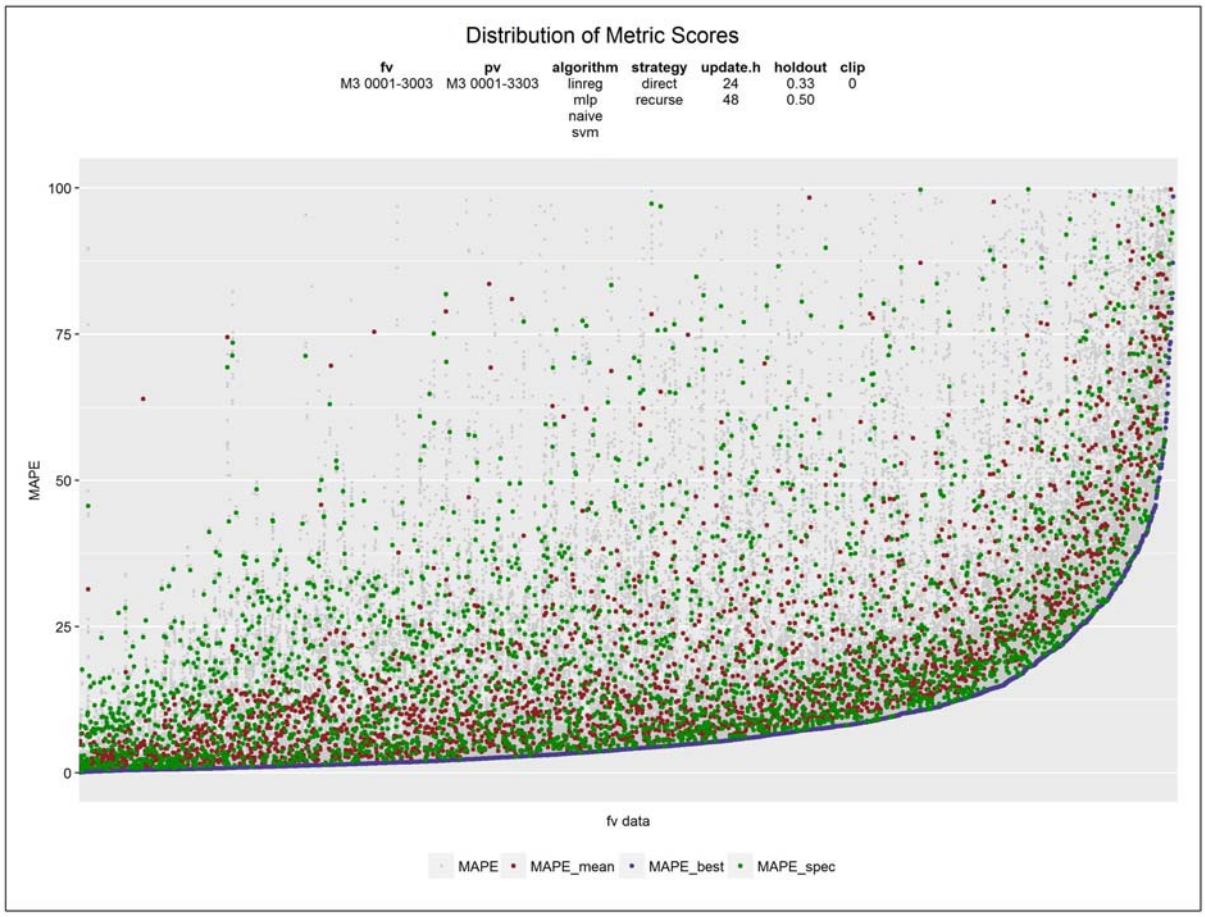


Figure 2-11: Distribution of metric scores across data sources. Data sources from the M3 Forecasting Competition. 3,003 data sources x 32 techniques = 96,096 forecasts and evaluations. Data sources are arranged along the x-axis. *Violet* is best MAPE score among 32 techniques applied to one data source. *Crimson* is mean MAPE score among 32 techniques applied to one data source. *Green* is MAPE score for one specific technique applied to one data source.

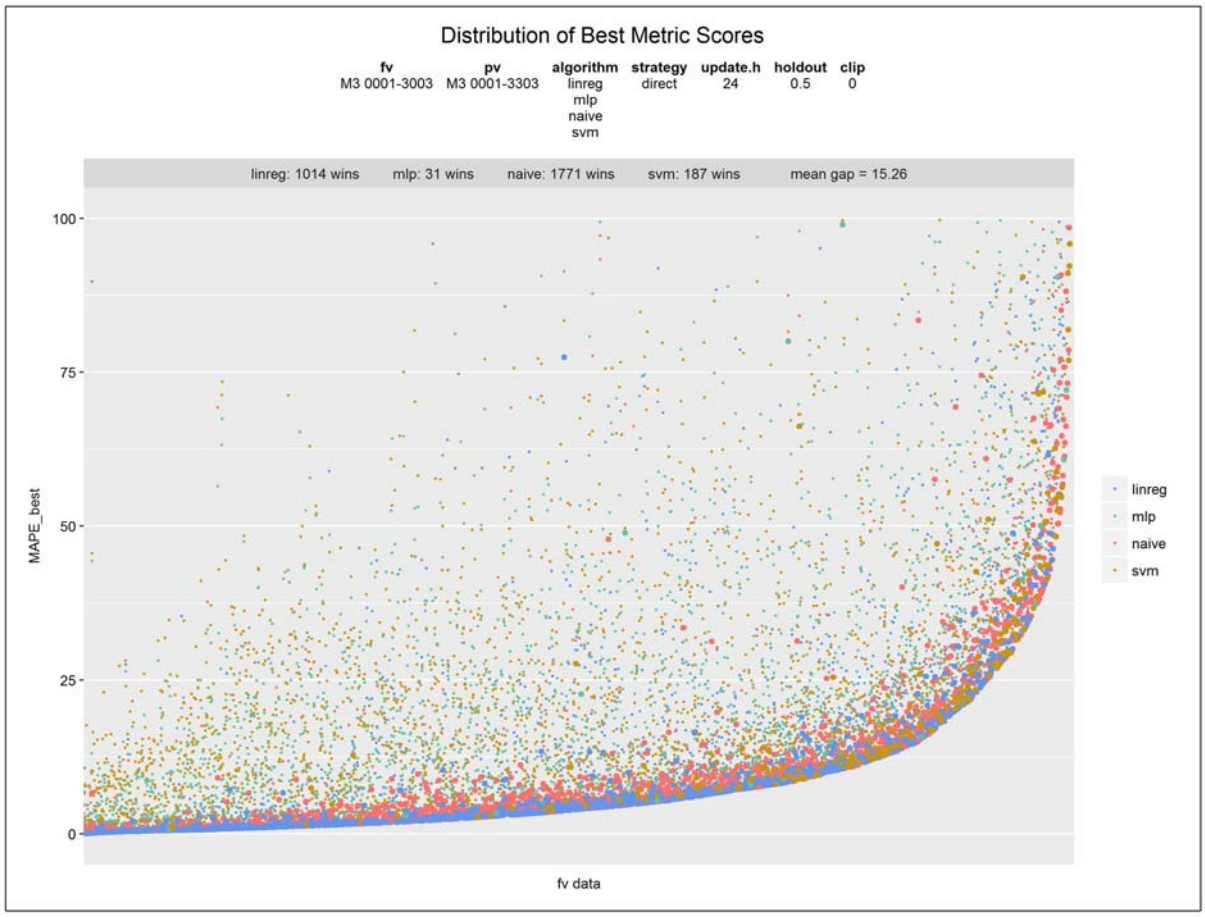


Figure 2-12: Distribution of best metric scores across data sources, by algorithm class, 4 techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources x 4 techniques = 12,012 forecasts and evaluations. Data sources are arranged along the x-axis. *Blue* is MAPE score of linear regression-based technique applied to one data source. *Green* is MAPE score of perceptron-based technique applied to one data source. *Red* is MAPE score of naïve-based technique applied to one data source. *Gold* is MAPE score of support vector regression-based technique applied to one data source. *Large point* is best MAPE score among 4 techniques applied to one data source.

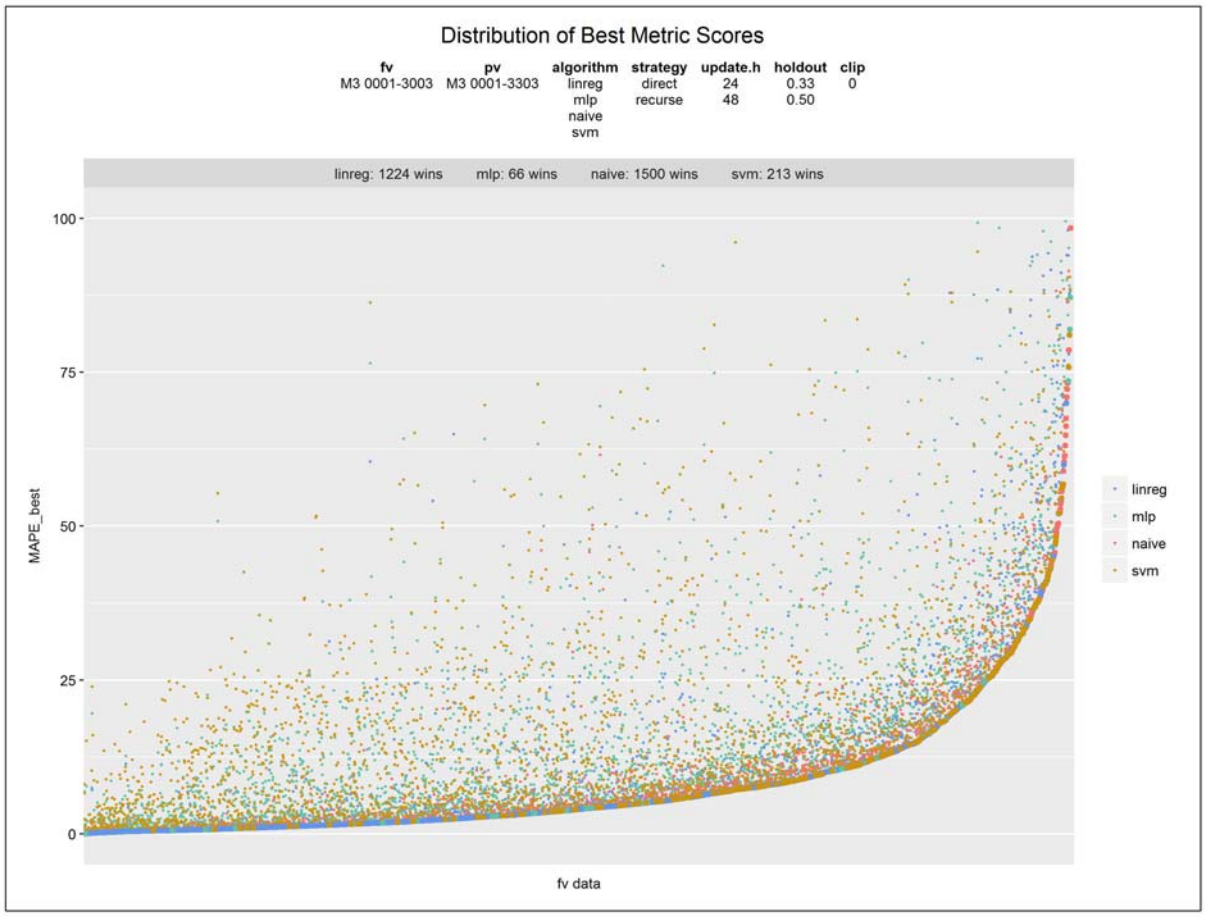


Figure 2-13: Distribution of best metric scores across data sources, by algorithm class, 32 techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources x 32 techniques = 96,096 forecasts and evaluations. Data sources are arranged along the x-axis. *Blue* is best MAPE score among 8 linear regression-based techniques applied to one data source. *Green* is best MAPE score among 8 multilayer perceptron-based techniques applied to one data source. *Red* is best MAPE score among 8 naïve-based techniques applied to one data source. *Gold* is best MAPE score among 8 support vector regression-based techniques applied to one data source. *Large point* is best MAPE score among 32 techniques applied to one data source.

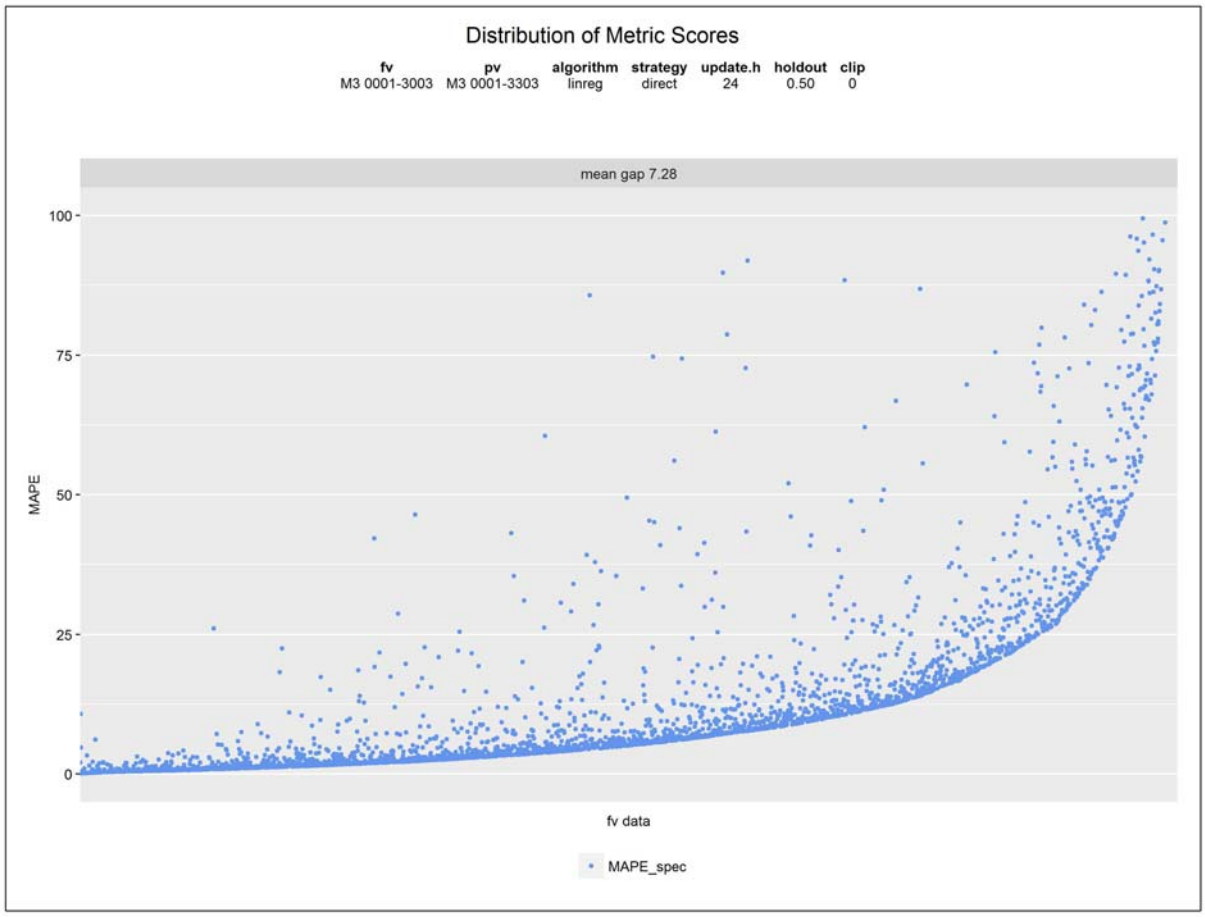


Figure 2-14: Distribution of metric scores across data sources, lock-in algorithm class, 1 technique. Data sources from the M3 Forecasting Competition. 3,003 data sources x 1 technique = 3,003 forecasts and evaluations. Data sources are arranged along the x-axis. *Blue* is MAPE score of linear regression-based technique applied to one data source.

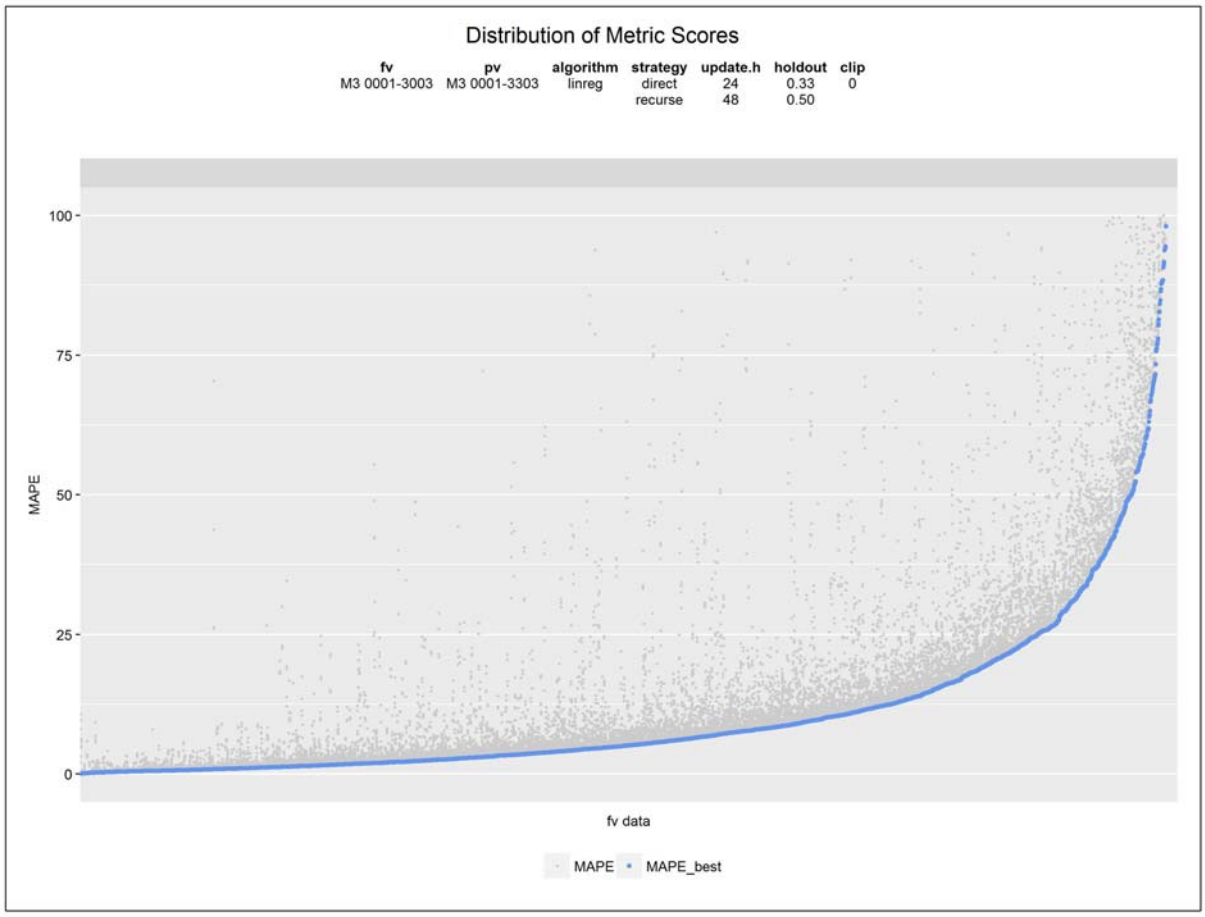


Figure 2-15: Distribution of best metric scores across data sources, lock-in algorithm class, 8 techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources x 8 techniques = 24,024 forecasts and evaluations. Data sources are arranged along the x-axis. *Blue* is best MAPE score among 8 linear regression-based techniques applied to one data source. *Large point* is best MAPE score among 8 techniques applied to one data source.

3 DECISIONS AND DATA ENTROPY AS A PREDICTOR OF FORECASTABILITY

"Tonight's forecast: dark. Broken up tomorrow by widely scattered light."

– George Carlin

3.1 Research Questions

In this chapter, we address the following research questions:

- How well does data entropy predict forecastability?
- How do forecasting process decisions affect the power of entropy to predict forecastability?

3.2 Research Approach

It would be useful to forecasting practitioners to know a priori something about how much a given reference series lends itself to being forecasted, so that commensurate effort can be planned for accordingly. Studies suggest that data entropy of reference series can predict means or bounds on associated forecasting performance in some cases, and so may be a useful tool in this regard. [148,149,150,151,152] We apply our analysis methods to expand on these studies, exposing the effects of varying several decisions on the relationship between entropy and forecastability, and compare our results. In our version, forecastability is measured first with respect to the set of techniques reflected by the decisions, and then with respect to the set of best performing techniques only, rather than with respect to a single technique as is common in the other studies.

3.3 Scope of Analysis

We construct and explore performance of 96,096 forecasts –32 techniques across 3,003 data sources – by varying data source and decisions for algorithm class, extension rule, update cycle, and holdout. The data sources are those used in the M3 Forecasting Competition. [156] MASE is locked-in as the metric decision. The entropy function assumes a default parameterization (similarity criterion r = 0.2 x standard deviation of data) applied to all data.

Reference data source	3,003 options	From M3 Forecasting Competition. Any of 3,003 anonymous time series datasets of length 20 to 144 time steps.
Predictor data source	3,003 options	*Same as reference data source.*
Algorithm class	4 options	linear regression, multilayer perceptron, naïve, support vector regression
Extension rule	2 options	direct, recurse
Update cycle	2 options	1 time step, 2 time steps
Holdout	2 options	33%, 50%
Metric	lock-in	MASE (based on training data mean)

3.4 Results

In the context of many decisions, but looking at one technique at a time, correlation between entropy and forecastability across reference series ranges from $R^2 = 0.00$ to $R^2 = 0.54$ (Figure 3-1). Only techniques that use the naïve algorithm class correlate at $R^2 > 0.12$. Most are much worse, e.g., one technique that uses support vector regression correlates at 0.11 (Figure 3-2).

Across all techniques, correlation between entropy and forecastability is almost completely absent (Figure 3-3).

Across only the best performing techniques for each reference series, correlation between entropy and corresponding (best) forecast MASE scores is much higher: $R^2 = 0.63$ (Figure 3-4). Twenty-eight of 32 techniques perform better than all others on at least one data source. The algorithm class decision appears to play a role in determining which techniques do better than others (Figure 3-5). Linear regression-based techniques correlate at $R^2 = 0.69$ (Figure 3-6), multilayer perceptron-based techniques correlate at $R^2 = 0.61$ (Figure 3-7), naïve-based techniques correlate at $R^2 = 0.55$ (Figure 3-8), and support vector regression-based techniques correlate at $R^2 = 0.24$ (Figure 3-9). Further, the least correlated techniques are not well-represented among the best performing techniques.

The span decision (in these experiments determined by the reference series) also appears to play a role. Across only the best performing techniques for each reference series at span ≤ 36 time steps, correlation is low: $R^2 = 0.13$ (Figure 3-10). As span increases to 37-72 time steps, correlation increases to $R^2 = 0.65$ (Figure 3-11). Spans of 73-108 time steps and ≥ 109 times steps correlate at $R^2 = 0.73$ (Figure 3-12, Figure 3-13). A clear pattern is evident suggesting that the predictive power of entropy increases asymptotically to about R^2=0.73 as span increases (Figure 3-14). Further, as the span and predictive power increase, so does the sensitivity of forecastability to entropy (Equation 3-1).

Equation 3-1:
Empirical relationship between entropy and forecastability,
based on M3 Forecasting Competition data at long span.

$$\frac{1}{F(D,d)} \equiv MASE_{best}(D,d) \cong 0.35 \times e(D) - 0.04$$

Where …
F is measure of forecastability,
D is dataset (used as reference and predictor series),
d is vector of decisions,
e is entropy function,
$MASE_{best}$ is mean absolute scaled error of forecast using best technique.

3.5 Comparison to Benchmark Studies

Catt investigated entropy as a potential predictor of forecastability across the M3 Forecasting Competition 3,003 time series. [148] In the study, forecastability is measured in terms of MASE, where training data mean is assumed as the default forecast. In our parlance, the study varies the data source among these time series. All other decisions are locked-in, but for the caveat that holdout is specified as a fixed number of time steps, not as a fraction of the total number of time steps – and depends on the data source. Also, the entropy function is parameterized (by setting similarity criterion r) to optimize the correlation between the entropy of training data and the corresponding MASE scores for forecasts made on the training data. Entropy is calculated for training data only.

Reference data source	3,003 options	From M3 Forecasting Competition. Any of 3,003 anonymous time series datasets of length 20 to 144 time steps.
Predictor data source	3,003 options	*Same as reference data source.*
Algorithm class	lock-in	Holt linear trend
Extension rule	lock-in	direct
Update cycle	lock-in	1 time step
Holdout	lock-in	6, 8, or 18 time steps, depending on data source
Metric	lock-in	MASE (based on training data mean)

The study finds that entropy defined in terms of a specific technique is modestly correlated with forecastability at R^2=0.56, about the same as in our study for forecastability defined in terms of the best correlated technique.

3.6 Insights

In practice, forecastability as measured assuming many decisions at play may be more useful than as measured assuming a single technique, because it avoids the need to treat a specific technique as a proxy for the best performing technique that will ultimately be selected by the forecasting practitioner. Happily, we see that accounting for the set of best techniques actually unleashes the power of entropy to better predict forecastability.

We glean the following insights from our results, with the requisite caveat that they are based on one specific set of data and a practically scoped set of experiments.

> For forecastability defined in terms of a specific technique, entropy is a weak to modest predictor of forecastability, depending on the metric and technique decisions.

> For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, entropy is a strong predictor of forecastability, more or less so depending on the technique decisions for algorithm class and span.

> For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, the predictive power of entropy grows asymptotically along with decisions to increase span.

> For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, the sensitivity of forecastability to entropy grows along with decisions to increase span.

3.7 Tables and Data Visualizations

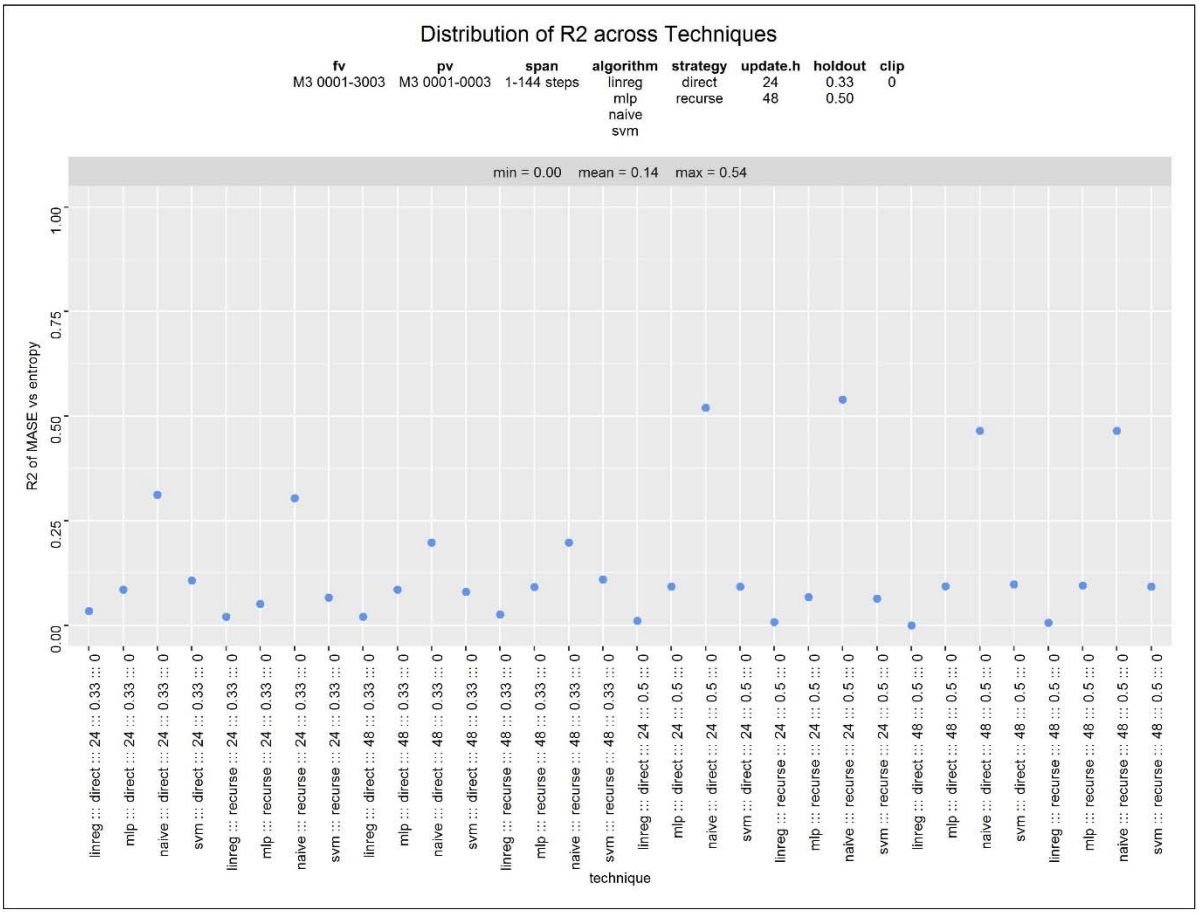


Figure 3-1: Distribution of correlations of metric score-to-entropy across techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Techniques are arranged along the x-axis.

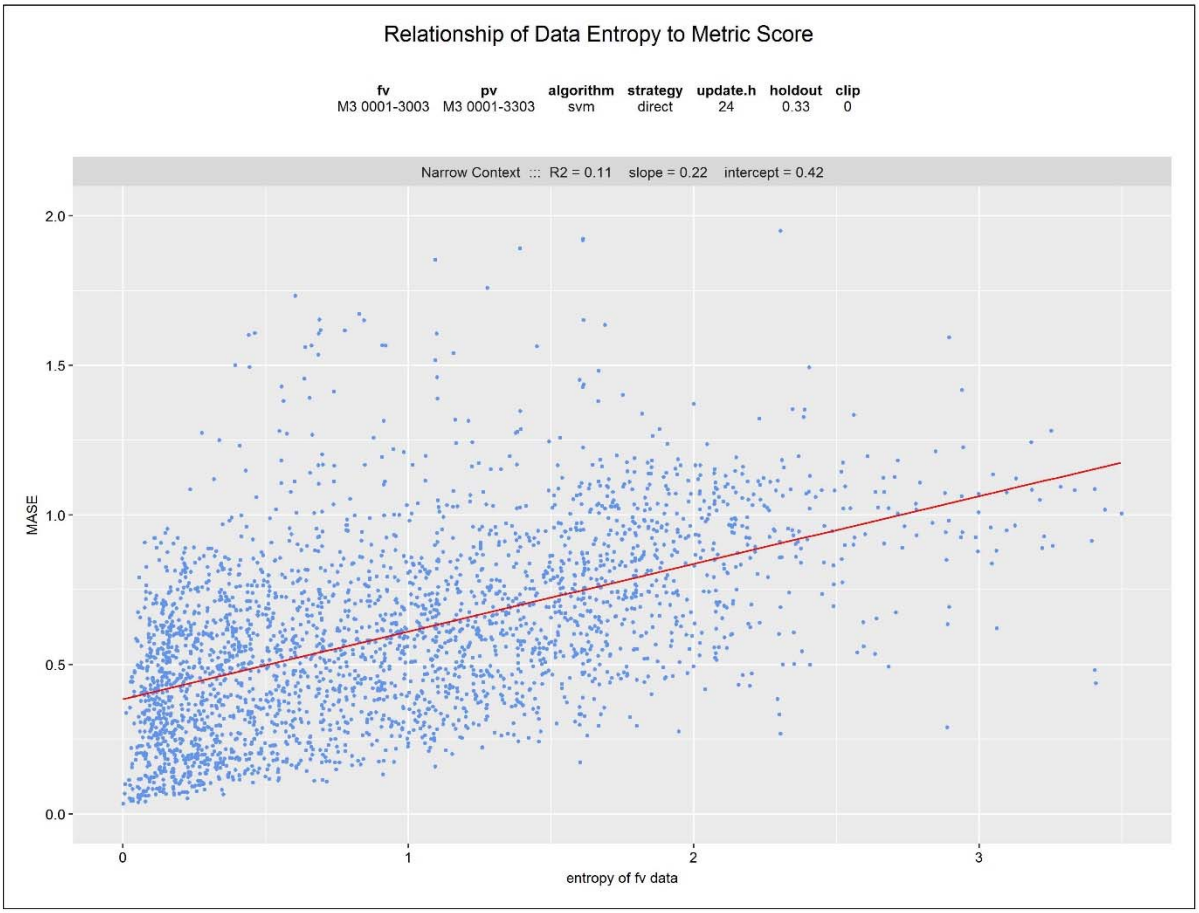


Figure 3-2: Metric score vs. entropy, 1 technique. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 1 technique = 3,003 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis.

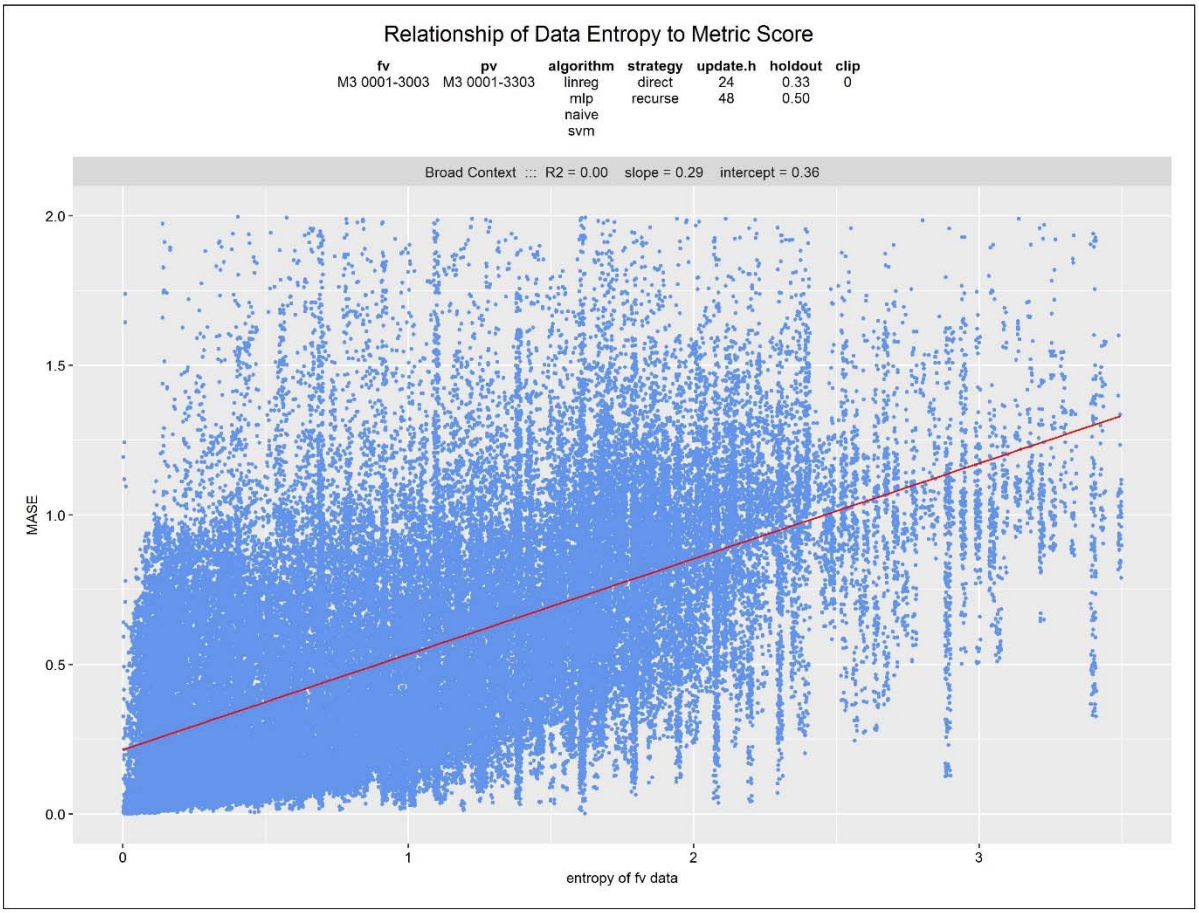


Figure 3-3: Metric score vs. entropy, 32 techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis.

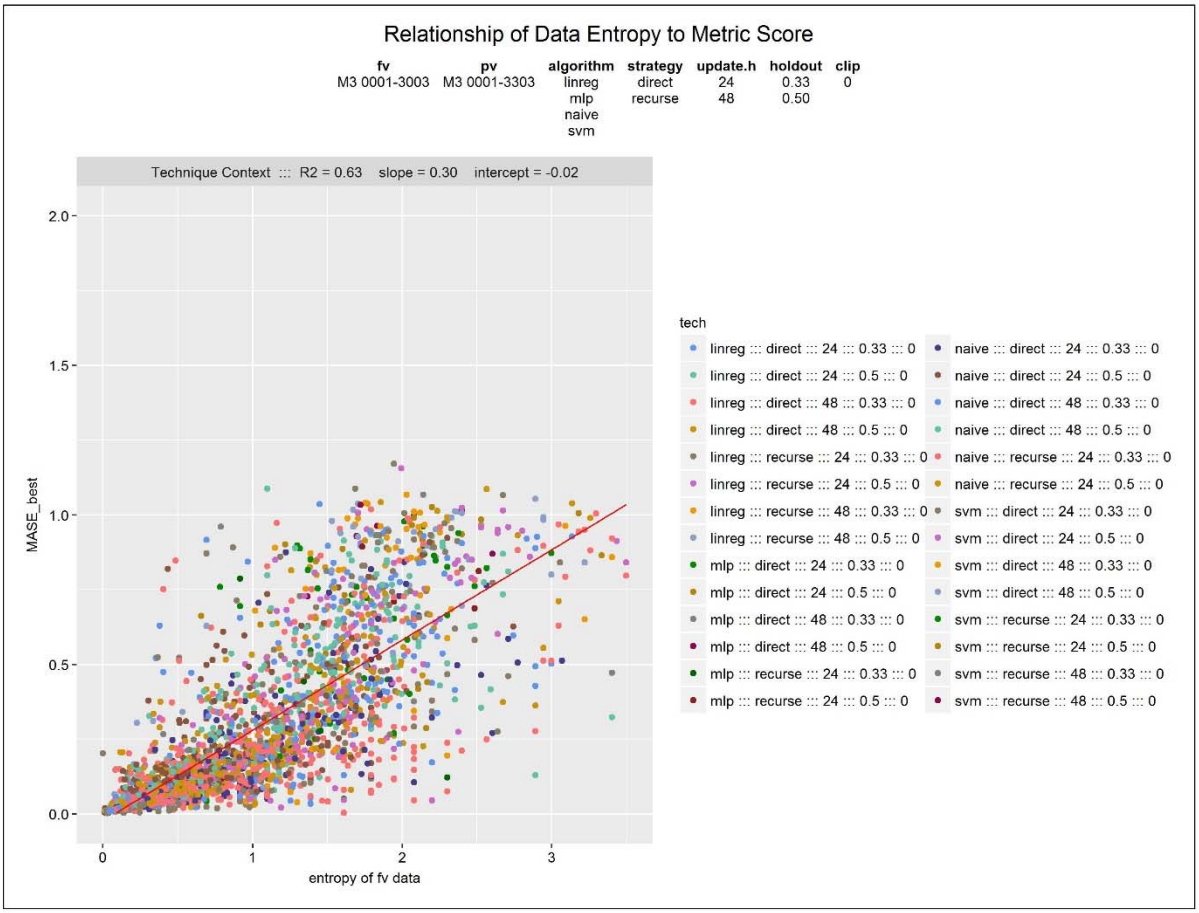


Figure 3-4: Best metric score vs. entropy, 32 techniques. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are best technique applied to one data source.

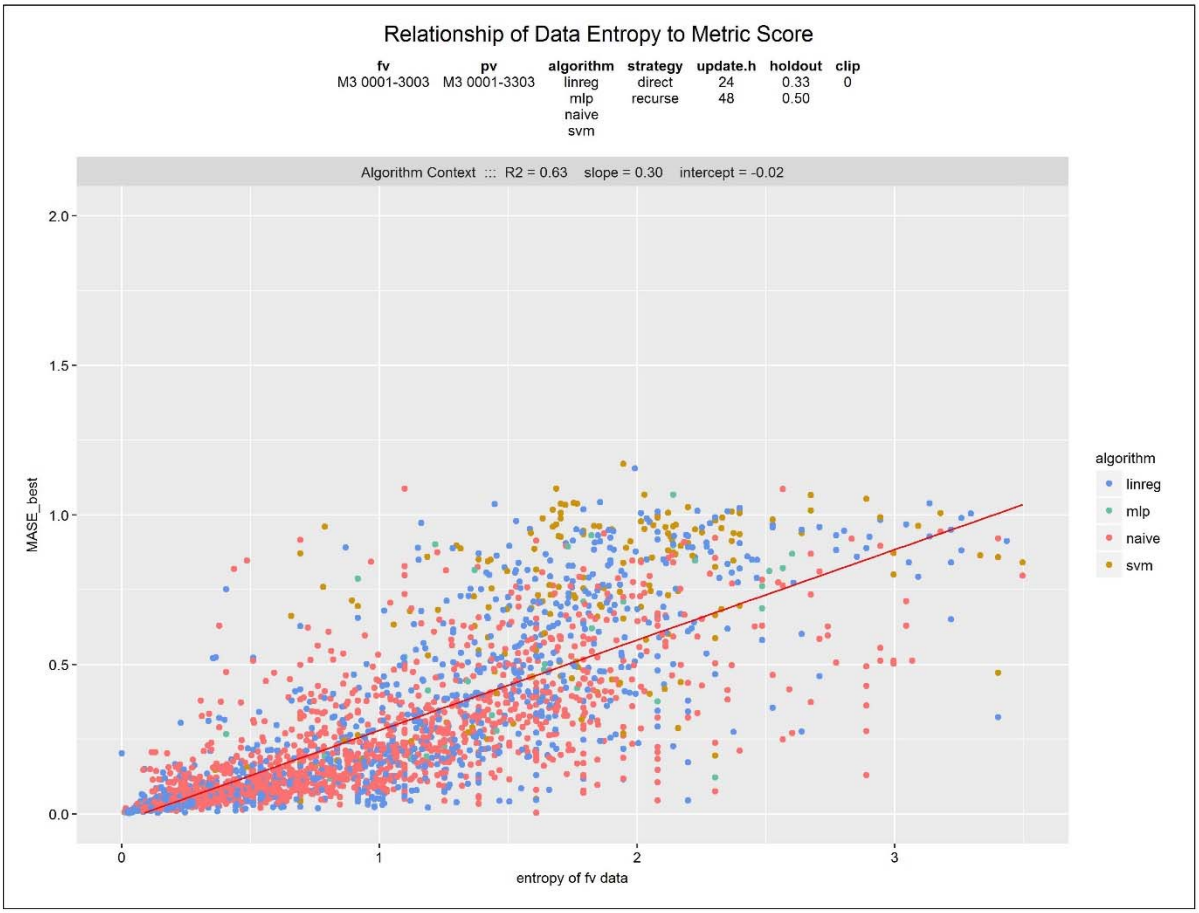


Figure 3-5: Best metric score vs. entropy, by algorithm class. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are algorithm class of best technique applied to one data source.

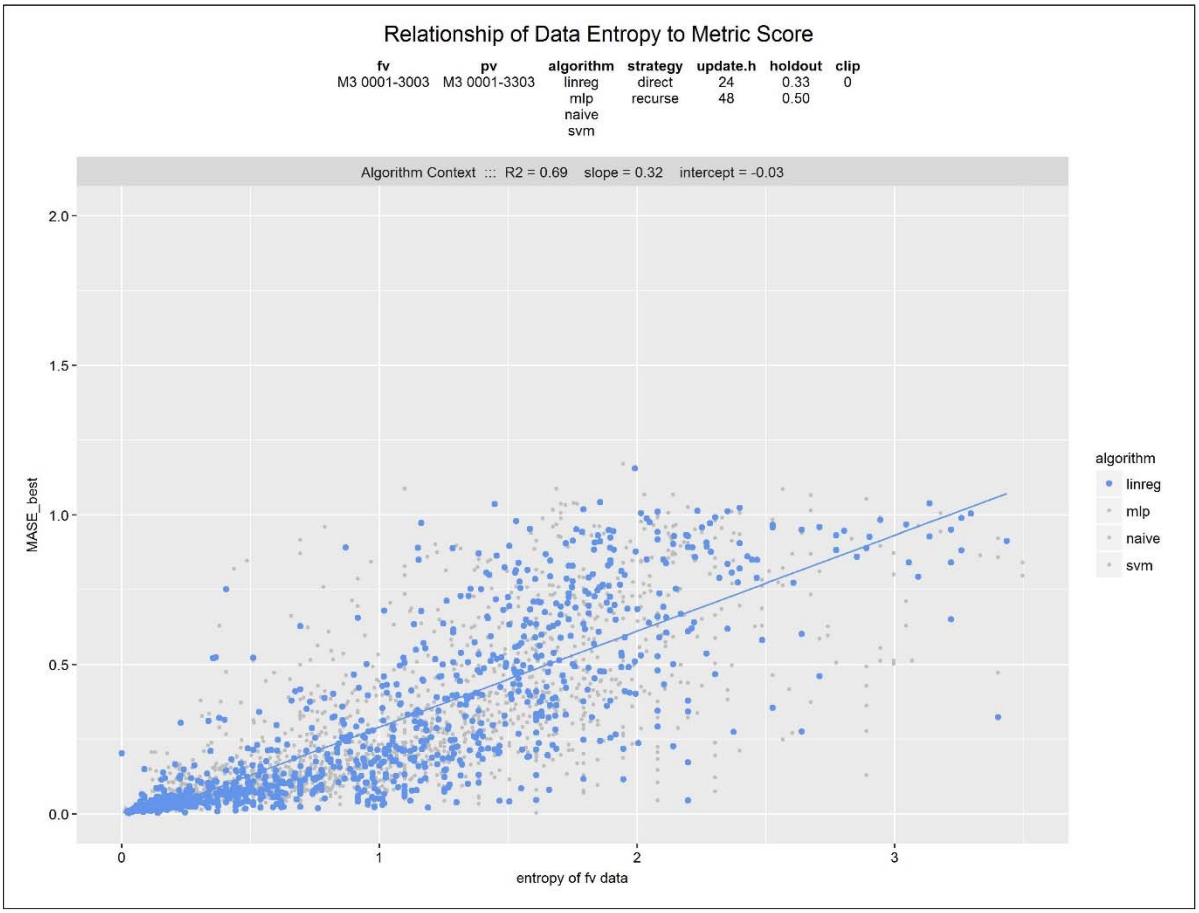


Figure 3-6: Best metric score vs. entropy, linear regression. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Blue* is best technique applied to one data source, filtered for linear regression algorithm class.

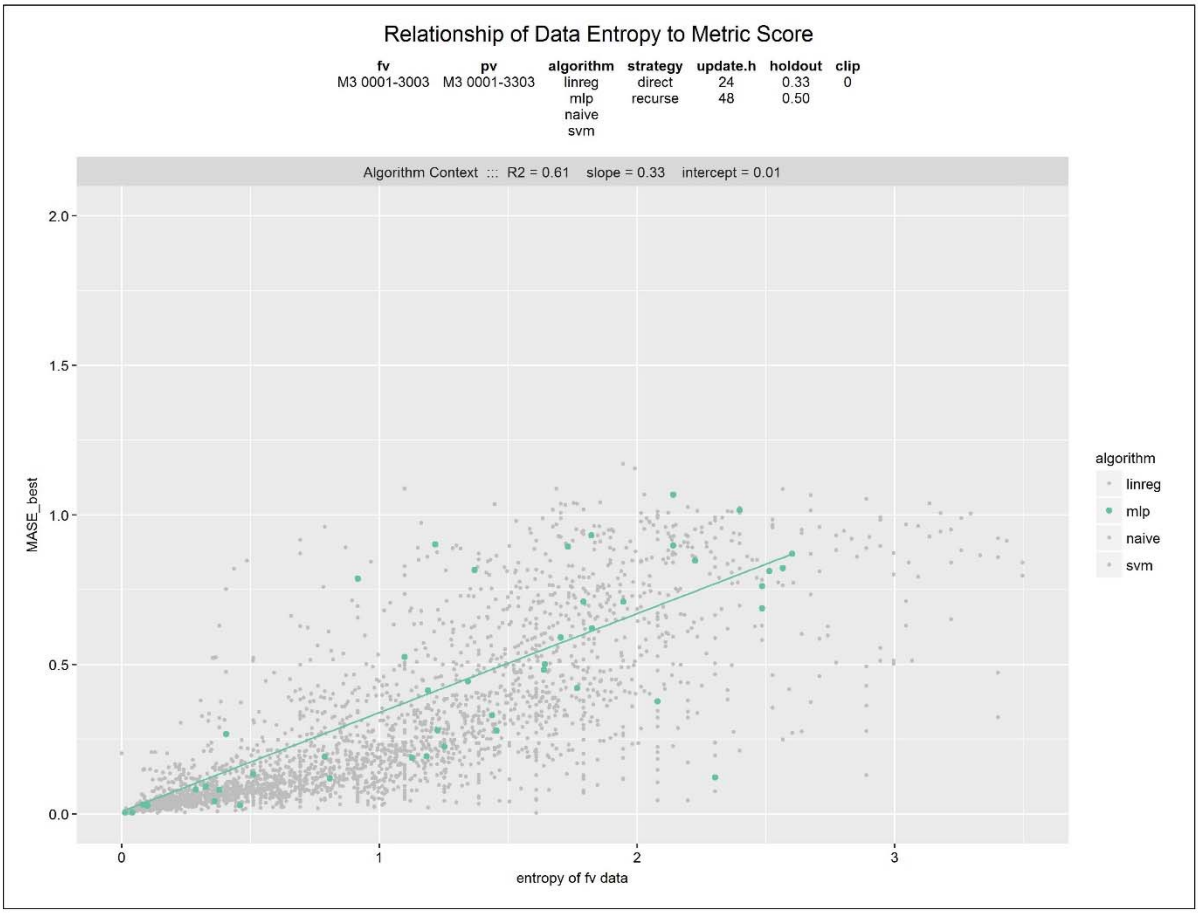


Figure 3-7: Best metric score vs. entropy, multilayer perceptron. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Green* is best technique applied to one data source, filtered for multilayer perceptron algorithm class.

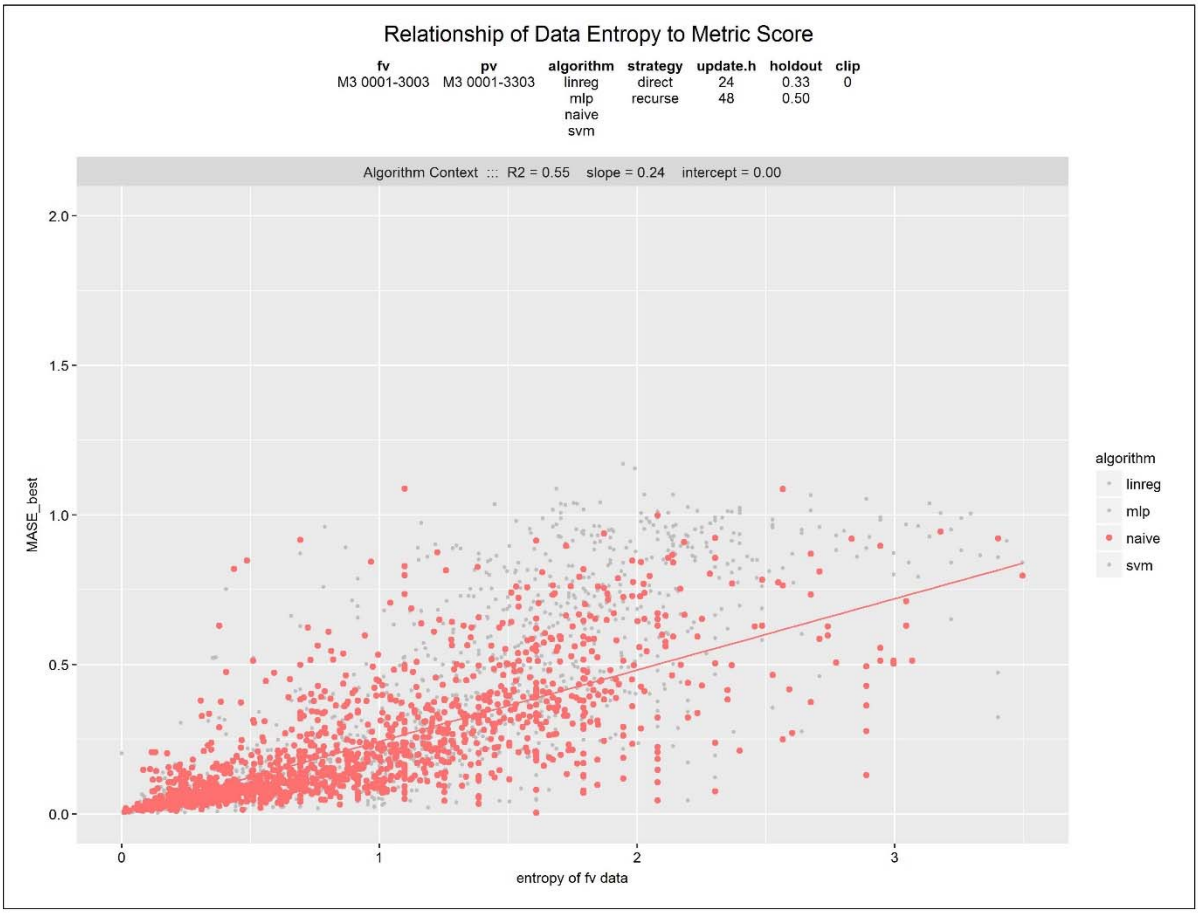


Figure 3-8: Best metric score vs. entropy, naive. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Red* is best technique applied to one data source, filtered for naive algorithm class.

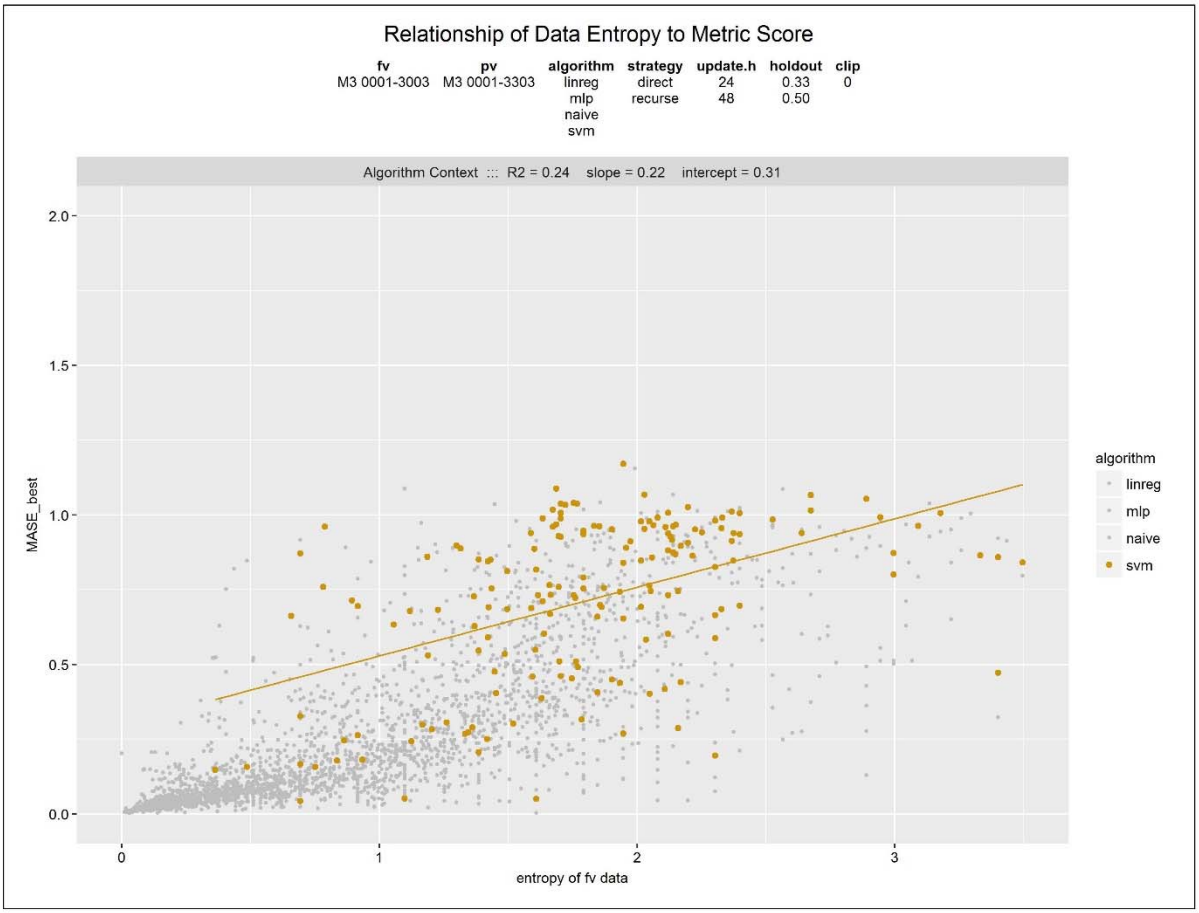


Figure 3-9: Best metric score vs. entropy, support vector regression. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Gold* is best technique applied to one data source, filtered for support vector regression algorithm class.

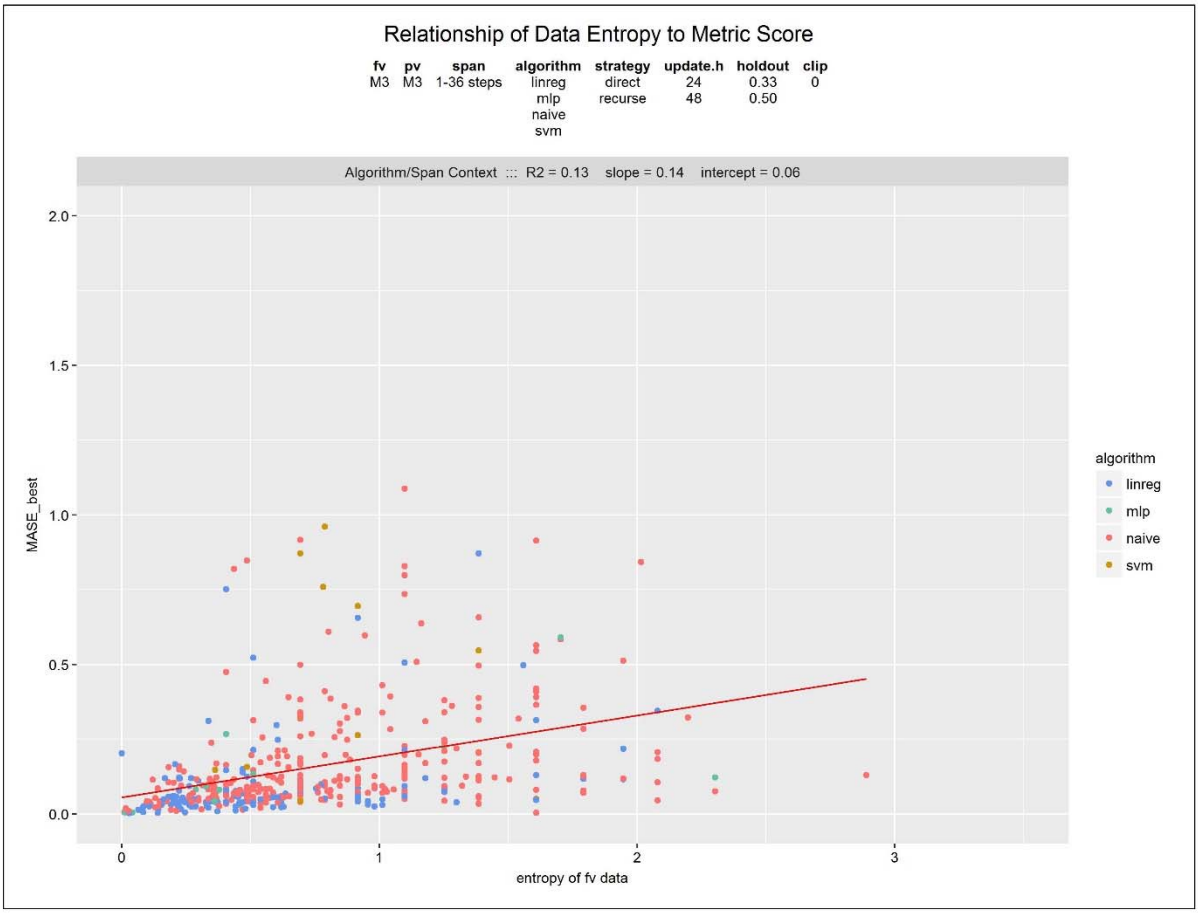


Figure 3-10: Best metric score vs. entropy, short span. Data sources of length 1-36 time steps from the M3 Forecasting Competition. 32 techniques. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are algorithm class of best technique applied to one data source.

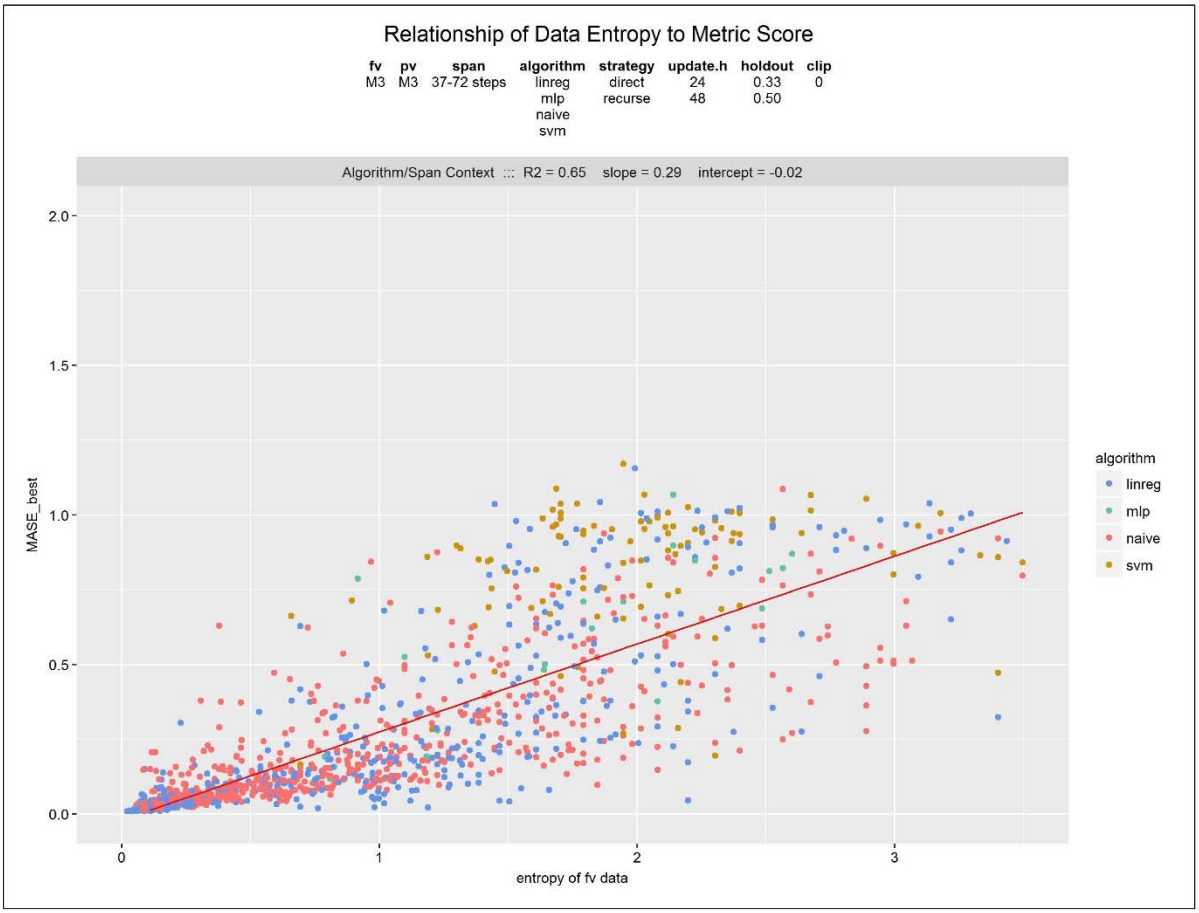


Figure 3-11: Best metric score vs. entropy, medium span. Data sources of length 37-72 time steps from the M3 Forecasting Competition. 32 techniques. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are algorithm class of best technique applied to one data source.

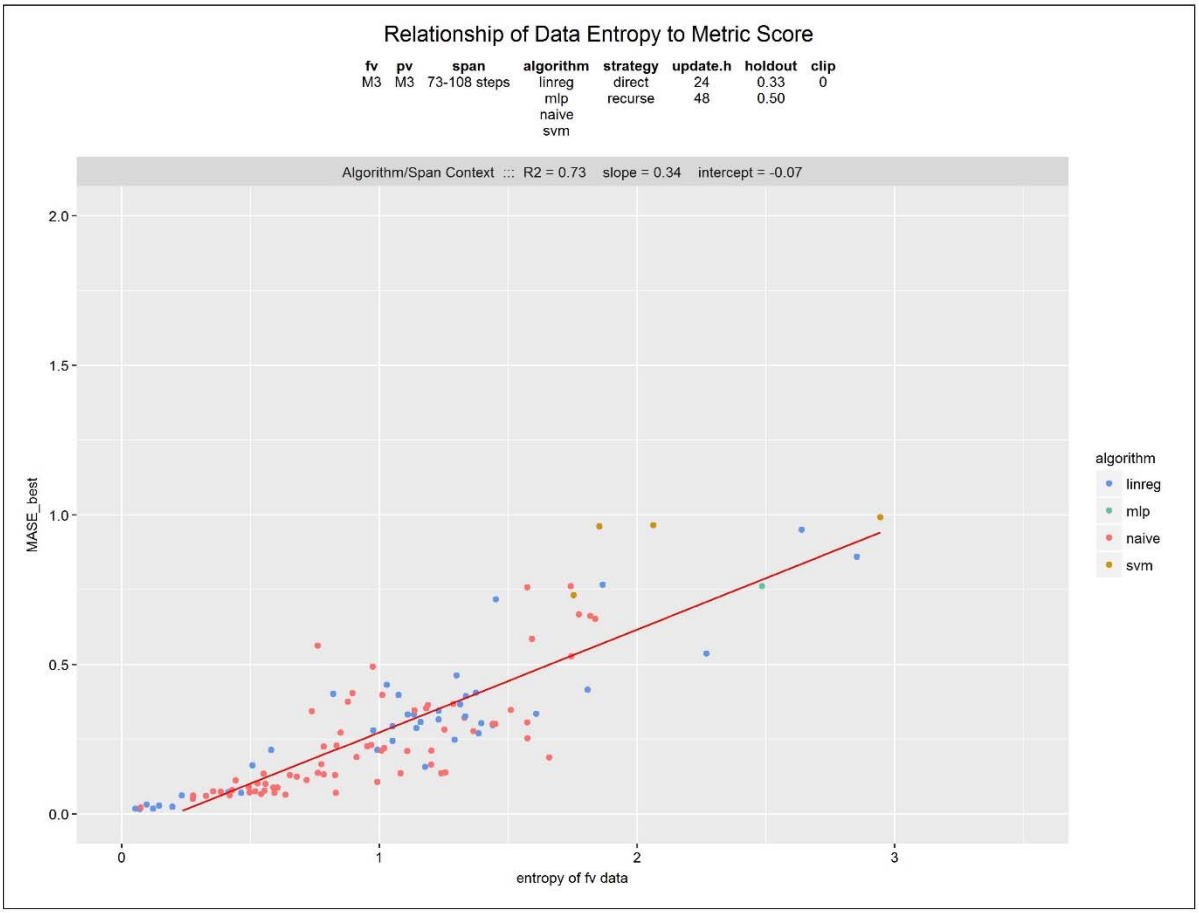


Figure 3-12: Best metric score vs. entropy, medium-long span. Data sources of length 73-108 time steps from the M3 Forecasting Competition. 32 techniques. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are algorithm class of best technique applied to one data source.

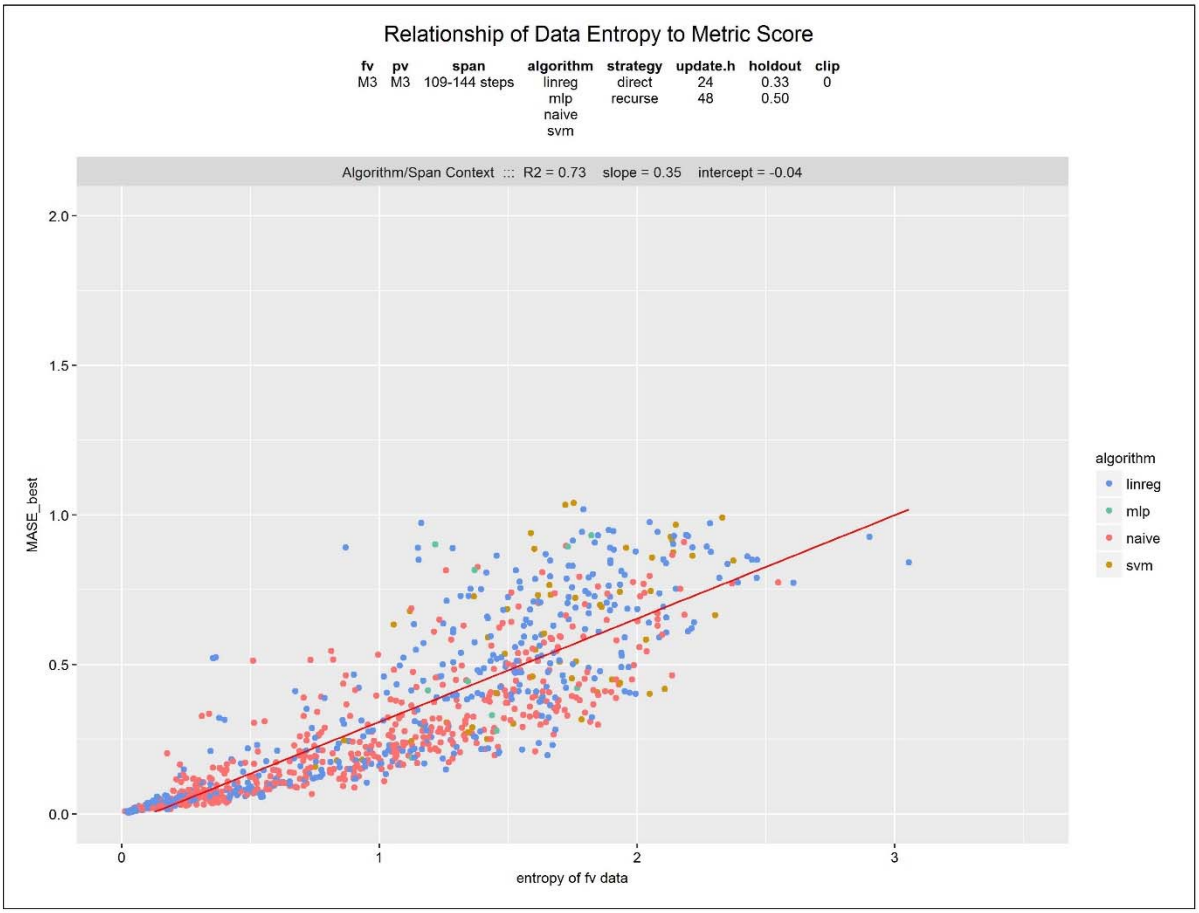


Figure 3-13: Best metric score vs. entropy, long span. Data sources of length 109-144 time steps from the M3 Forecasting Competition. 32 techniques. Evaluation by MASE score. Entropy values are arranged along the x-axis. *Colors* are algorithm class of best technique applied to one data source.

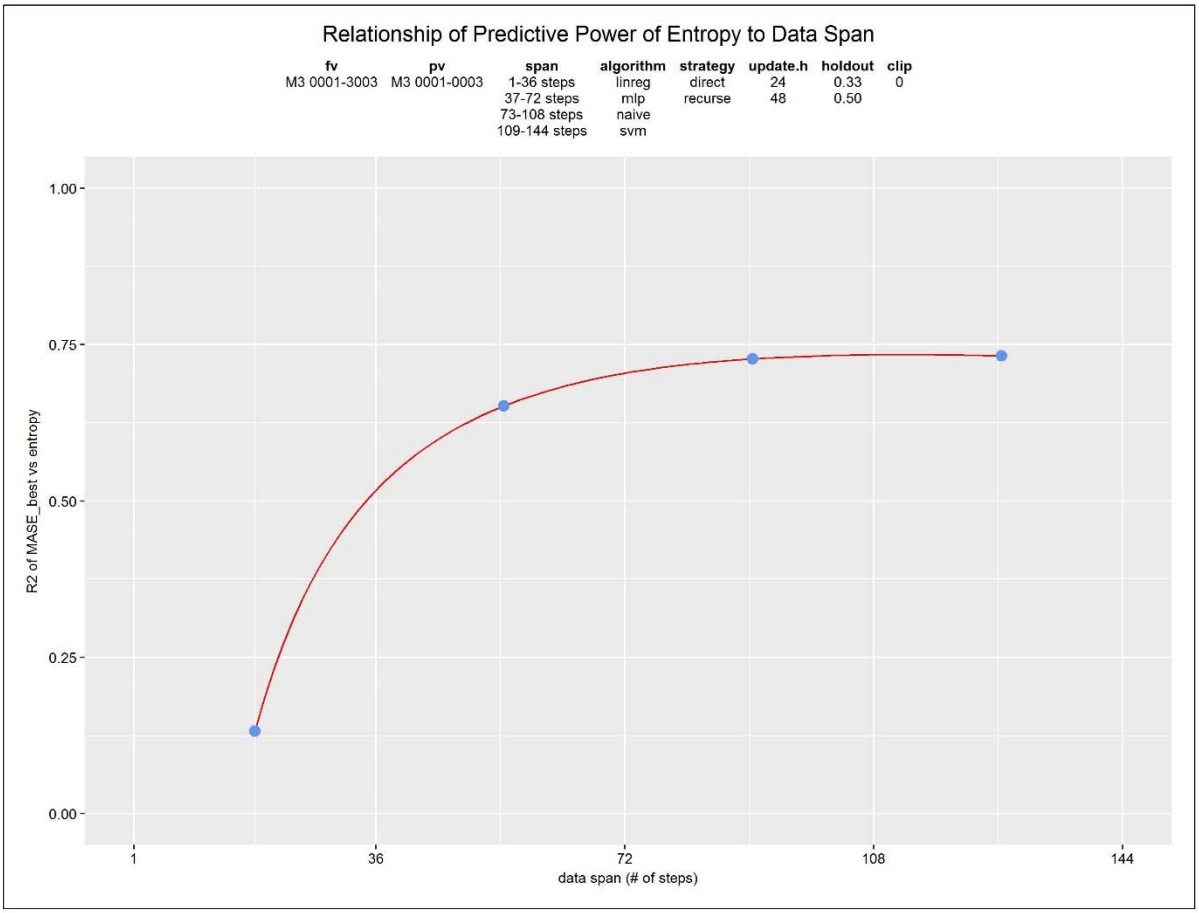


Figure 3-14: Correlation of metric score-to-entropy vs. data span. Data sources from the M3 Forecasting Competition. 3,003 data sources from x 32 techniques = 96,096 forecasts and evaluations. Evaluation by MASE score. Data spans are arranged along the x-axis.

4 TECHNIQUE DECISIONS AND RESIDENTIAL ELECTRICITY DEMAND ESTIMATION

"If you can look into the seeds of time, and say which grain will grow and which will not, speak then unto me."

– William Shakespeare

4.1 Research Questions

In this chapter, we address the following research questions:

- How well do residential electricity demand forecasting techniques perform?
- How can smart electric grid data be utilized to improve residential electricity demand forecasting performance?

4.2 Literature Review

Electricity demand forecasting research has been quadrifurcated based on look-ahead and update cycle: very short-term, short-term, medium-term, and long-term. [14] By look-ahead, we mean how far into the future a forecast begins. By update cycle, we mean how far from the time a forecast begins does it continue before a new forecast is allowed to be made. The term horizon is sometimes used interchangeably with look-ahead or update cycle, so here we use only the latter terms to avoid confusion. Challenges, opportunities, and benefits of the different look-ahead ranges have proven specialized enough to warrant unique study approaches for each range. Various definitions for the four ranges have been proposed to better organize the relevant research and application in power utilities (Table 2-1, Table 2-2). [14,18,23]

Day-ahead to week-ahead is the range typically considered by electricity demand forecasting practitioners when scheduling electricity generation and transmission, and as such has motivated investigations specific to these look-aheads.

Many electricity demand forecasting studies score forecasting performance per mean absolute percentage error (MAPE) or root mean square error (RMSE). MAPE of 4% seems to be a utility industry de facto standard for the threshold qualifying a forecast as practical.

Statistical algorithms have been applied to electricity demand forecasting since at least 1966. [2,76] Machine learning has been applied to electricity demand forecasting since the 1990s. [42,76] Machine learning algorithms based on neural networks, including multilayer perceptrons, and support vector regression in particular showed early promise for electricity demand forecasting and continue to account for much research. [42,77] A form of neural network coined "extreme learning" has generated controversy around its originality, but is reported especially effective in electricity demand forecasting. [30]

With the recent advent of smart electric grids, electricity usage data measured at fine temporal and geographic resolutions have become available to researchers for the first time. Notable smart electric grid datasets include those from the Ireland Commission for Energy Regulation (Smart Metering Project in Dublin, Ireland), the Australian Government (Smart Grid / Smart City Customer Trial in Australia), Pacific Gas and Electric (Energy Data Request Program in California, United States), and the University of Texas at Austin (Pecan Street Demonstration in Texas, United States). [153,155,157,158]

With more data detail comes new decisions around how much data to use, how to sample it, how to cluster it, how to temporally magnify it, and how to aggregate it.

The Global Energy Forecasting Competition (GEFCom), a competition focusing specifically on electricity demand-related forecasting, has been held twice, in 2012 and 2014. [137,138,139,140,142,143] The winning technique in 2012 employed a version of the multiple

linear regression algorithm. The winning technique in 2014 employed a quantile generalized additive model algorithm.

4.3 Research Approach

We apply our analysis methods and computation platform to study 8,016 week-ahead and day-ahead electricity demand forecasting techniques scored by 60 metrics, utilizing smart electric grid data on 782 households collected by the Ireland Commission for Energy Regulation Smart Metering Project in 2009-10. The techniques incorporate nearest neighbor, linear regression, multilayer perceptron, naïve, support vector regression, and decision tree algorithms.

4.4 Scope of Analysis

We scope our analysis to the effects of 12 decisions and a collection of electricity-related datasets, locking in 5 of the decisions and varying the remaining 7 decisions over a range of options. We choose these particular decisions because they are represented in several other studies that consider their effects in isolation, whereas we are interested in their combined effect. We choose this particular collection of datasets because it reflects the finest granularity, multi-year electricity usage data now publicly available.

We are interested in both week-ahead and day-ahead forecasting settings, so we instantiate two versions of the model – one for week-ahead and one for day-ahead – and conduct the analysis in two parts accordingly, reporting the results separately.

4.4.1 Model Instantiation for Week-Ahead Forecasting

For the week-ahead part of our analysis, we lock-in a one-day time step size. Our model instantiations then vary across options for seven other decisions, identifying 2,880 techniques, all scored by 60 metrics, for a total of 172,800 vectors = 5 update cycle options x 3 span options x 6 algorithm class options x 2 extension rule options x 4 holdout options x 4 clip options x 60 metrics. We vary these particular decisions to highlight the potential impact of small variations

in choices around techniques that are commonly based on tacit assumptions about their importance in other studies.

The algorithm class options comprise representatives from five popular, distinct forecasting approaches, plus a naïve algorithm that serves as a benchmark. The naïve algorithm used here always forecasts no change in electricity demand at one-week ahead. The multilayer perceptron and support vector regression algorithms are statically tuned to hyper-parameter values that work well for one typical technique.

Objective Decisions	lock-in	Reference data source	Ireland CER electricity usage
	lock-in	Reference look-ahead	1 week
	5 options	Update cycle	1 day 2 days 3 days 5 days 7 days
	lock-in	Time step size	1 day
Data Strategy Decisions	3 options	Span	6 months 12 months 18 months
Form Decisions	6 options	Algorithm class	k-nearest neighbor linear regression multilayer perceptron naïve support vector regression decision tree
	lock-in	Predictor data sources	Ireland CER electricity usage + WU temperature + day of week
	lock-in	Predictor look-backs	(0, 1 week, 2 weeks, 3 weeks) + (0) + (0)

	2 options	Extension rule	direct recurse
Training and Testing Decisions	4 options	Holdout %	20% 30% 40% 50%
	4 options	Clip %	0 1% 10% 20%
Metric and Penalty Decisions	60 options	Metric function or penalty function	*any of 30 metrics functions or 30 penalty functions*

4.4.2 Model Instantiation for Day-Ahead Forecasting

For the day-ahead part of our analysis, we lock-in a one-hour time step size. Our model instantiations then vary across options for seven other decisions, identifying 648 techniques, all scored by 60 metrics, for a total of 38,880 vectors = 3 update cycle options x 3 span options x 4 algorithm class options x 2 extension rule options x 3 holdout options x 3 clip options x 60 metric options.

The four algorithm class options have appeared in other day-ahead forecasting studies using the same reference data source.

Objective Decisions	lock-in	Reference data source	Ireland CER electricity usage
	lock-in	Reference look-ahead	1 day
	3 options	Update cycle	1 hour 12 hours 24 hours
	lock-in	Time step size	1 hour

Data Strategy Decisions	3 options	Span	6 months 12 months 18 months
Form Decisions	4 options	Algorithm class	linear regression multilayer perceptron naïve support vector regression
	lock-in	Predictor data sources	Ireland CER electricity usage + WU temperature + day of week + hour of day
	lock-in	Predictor look-backs	(0, 1 day, 2 days, 3 days, 4 days, 5 days, 6 days) + (0) + (0) + (0)
	2 options	Extension rule	direct recurse
Training and Testing Decisions	3 options	Holdout %	20% 33% 50%
	3 options	Clip %	0 10% 20%
Metric and Penalty Decisions	60 options	Metric function or penalty function	*any of 30 metrics functions or 30 penalty functions*

4.4.3 Data Sources

Our techniques all assume the same reference and predictor series, renderings of data collected by a real smart electric grid, pre-processed as necessary to prepare it for use by the platform.

For our reference series, we use electricity usage time series data from the Commission for Energy Regulation (CER) Smart Metering Project conducted by the Government of Ireland, assuming that demand exactly equaled usage in the absence of physical constraints or financial disincentives. [128,155] This smart electricity grid covered 7,444 businesses and households in and around Dublin, Ireland, measuring usage at that geographic granularity at half-hour intervals, spanning July 14, 2009 through December 31, 2010. Of these, we restrict our series to reflect 782 households that had no missing data throughout the entire period, and that were not affected by any experimental pricing policies. To consolidate the 782 time series, we convert to hourly resolution and sum across households at each time step, resulting in a 12,864-step time series. Hourly values for the aggregated population range from 245 kWh to 2,176 kWh, with mean 759 kWh.

Our predictor series (three for week-ahead, four for day-ahead) are electricity usage, temperature, day of week, and hour of day. For electricity usage, we use the same CER data. For temperature, we use actual temperature time series data reported for Dublin, Ireland, at hourly intervals, from the IBM Weather Underground weather data service. [159] For day of week and hour of day, we use time series data constructed per an integer coding scheme.

Reference series	Electricity usage. From Ireland Commission for Energy Regulation Smart Metering Project (CER). 30-minute intervals across 782 control households in Dublin, Ireland in 2009-10.
Predictor series	Electricity usage. From Ireland Commission for Energy Regulation Smart Metering Project (CER). Measured at 30-minute intervals across 782 households in Dublin, Ireland in 2009-10. *Same data source as for reference series.* Temperature. From IBM Weather Underground. Measured at 1-hour intervals at Dublin, Ireland in 2009-10. Day of week. Generated by analysis platform. Hour of day (for day-ahead only). Generated by analysis platform.

4.5 Results

For expediency and because MAPE is a de facto electricity forecasting standard, as noted earlier, we highlight results around MAPE-based techniques in our reports on sensitivity of metric scores, rank, and rank in context. Similarly, we highlight results around MAPE, MaxAE, and R2 in our reports on metric relationships and sensitivity of rank based on multiple metrics. Metrics like MaxAE and R2 are useful for capturing objectives that try to minimize the adverse effects of demand spikes, supply spikes, and alternating over- and under-supply. These three metrics in combination reflect important, but potentially conflicting, objectives.

4.5.1 Data Characterization

At one-day time step size, as used for our week-ahead forecasting, clearly evident are a negative correlation between the electricity usage and temperature series, annual seasonality in the electricity usage series, and significant correlation of electricity usage between any one time step and some multiple of exactly one day prior (Figure 4-3, Figure 4-4, Figure 4-5). The former correlation is well known and typical of cold climates, where households use electricity for heating, but less so for cooling. The same patterns are evident at one-hour time step size as used for our day-ahead forecasting.

Notably, two electricity usage spikes coincide with the Christmas 2009 and New Year 2010 holiday periods despite rises in temperature. A more pronounced spike coincides with the Christmas 2010 holiday period when temperatures fall to their coldest.

4.5.2 Sensitivity of Forecasts to Decisions

Evident in the distribution of forecasts is a tendency for many techniques to over-forecast a short time from the origin, and many more techniques to under-forecast more pronouncedly as the temporal distance from the origin increases (Figure 4-6, Figure 4-7). The pattern is evident in day-ahead forecasts and week-ahead forecasts.

4.5.3 Sensitivity of Metric Scores to Decisions

Twenty-five out of 30 metrics yield valid forecast performance evaluations for all forecasts. Metrics scores vary widely across the set of all techniques, as do metric scores across techniques that vary only in the metric used to evaluate forecast performance.

With week-ahead forecasts, different decisions influence forecast performance disproportionately (Figure 4-8). The span decision accounts for about 75% relative importance with respect to metric score across most of the metrics. Span remains highly influential even through the lens of a fixed algorithm decision, across all algorithm class options, though for a few metrics clip becomes more influential (Figure 4-9). With day-ahead forecasts, the decision of update cycle most influences metric score across all but one metric (Figure 4-10).

With week-ahead forecasts, MAPE scores are distributed in a distinctive pattern, ranging from 2.3% to 35.9%, with mean 10.8% (Figure 4-11). Scores are skewed toward the low end, but unevenly so, concentrated between 4% and 7%, with only 331 out of 2,880 scores falling below 4%. As noted earlier, MAPE of 4% is a de facto standard for the threshold qualifying a forecast as practical. Score distributions for other metrics from day-ahead forecasts and week-ahead forecasts line up in their own distinctive patterns.

With week-ahead forecasts, considered in the context of fixed algorithm class decisions, MAPE scores are usually best for techniques that use the naïve algorithm, all other decisions being equal, some by as little as 3 MAPE points and some by as much as 30 MAPE points (Figure 4-12). Techniques using support vector regression never beat their peers. In contrast, with day-ahead forecasts, techniques that use the naïve algorithm win less often, sharing the role of best among peers with techniques that use linear regression or support vector regression (Figure 4-13).

With week-ahead forecasts, looking deeper into the evident sensitivity of metric score to technique, most metric scores trend monotonically across span decision options, with techniques that use longer spans performing better (Figure 4-14). In contrast, no monotonic or

other regular trend appears across techniques that vary by holdout (Figure 4-15). For MAPE in particular, score either increases or decreases as holdout increases depending on the technique and the holdout option being compared (Figure 4-16). With day-ahead forecasts, the monotonicity seen in the span trends breaks down, with no clear relationship apparent between metric score and span decision (Figure 4-17).

4.5.4 Metric Relationships

With week-ahead forecasts, since only 25 of 30 metrics correspond to well-defined scores for all techniques, all other decisions being equal, we consider these 25 metrics and $\sum_{i=1}^{25-1} i = 300$ non-twin pairs of metrics as we look for relationships among them. Scores arranged by technique are strongly correlated at > 0.9 in 89 out of 300 pairs of metrics (Figure 4-18). Other pairs are modestly or poorly correlated. MAPE/MaxAE are correlated modestly at 0.66. MAPE/R2 are (negatively) modestly correlated at -0.68. MaxAE/R2 are modestly (negatively) correlated at -0.48. SAE is poorly correlated with most other metrics.

With day-ahead forecasts, all 30 metrics are in play, and lead to $\sum_{i=1}^{30-1} i = 435$ non-twin pairs. Most pairs are more strongly correlated when considering scores from day-ahead forecasts than from week-ahead forecasts (Figure 4-19).

4.5.5 Sensitivity of Ranks to Decisions

With week-ahead forecasts, techniques ranked per MAPE show a distinctive decision effectiveness signature (Figure 4-20). The top-ranked technique per MAPE uses a 12-month span, naïve algorithm, direct extension rule, 1-day update cycle, 50% holdout, and 20% clip (Figure 4-21, Table 4-1). Though this best performing technique uses the mid-length span, most of the 20 best performing techniques use the longest span (Figure 4-22). No one decision option is represented in all of the 20 best performing techniques, but 17 techniques use an 18-month span, and 17 (different) techniques use a 20% clip. None of the 20 best performing techniques use the multilayer perceptron or support vector regression algorithm, nor a clip less than 10%. Among the 20 best performing techniques, decisions about algorithm class,

extension rule, and holdout show more variation than do decisions about span, update cycle, and clip.

With day-ahead forecasts, techniques ranked per MAPE show a different decision effectiveness signature (Figure 4-23). The top ranked technique per MAPE uses an 18-month span, linear regression algorithm, direct extension rule, 1-hour update cycle, 33% holdout, and 20% clip (Figure 4-24, Table 4-2). Among the 20 best performing techniques, the decision for longest span still appears prominently, but not as much as with week-ahead forecasts (Figure 4-25). The 20 best performing techniques all use the linear regression algorithm, a direct extension rule, and a 1-hour update cycle. Among the 20 best performing techniques, no particular holdout or clip decision option is disproportionately represented.

With week-ahead forecasts, looking more closely at effects of individual decisions on MAPE score, the best 75% of techniques that use an 18-month span beat 80% of techniques that use a 12-month span and 94% of techniques that use a 6-month span. Further, while 100% of techniques that use an 18-month span beat only 15% of techniques that use a 12-month span, they beat 75% of techniques that use a 6-month span (Figure 4-26). Techniques that use the naïve algorithm mostly dominate other techniques, techniques that use support vector regression never dominate any other techniques (Figure 4-27). Techniques that use a recursive extension rule mostly dominate techniques that use a direct extension rule (Figure 4-28). Techniques that use shorter update cycles mostly dominate techniques that use longer update cycles, but not by much (Figure 4-29). Techniques that use the smallest holdout are dominated by the other techniques, but not by much (Figure 4-30). Techniques that use a larger clip mostly dominate techniques that use a smaller clip, but not by much (Figure 4-31).

With day-ahead forecasts, the span decision does not distinguish dominant techniques as it does with week-ahead forecasts (Figure 4-32). The best 13% of techniques that use linear regression beat all other techniques, but the rest of the techniques that use linear regression rarely dominate any other techniques (Figure 4-33). Techniques that use the shortest update cycle dramatically dominate the other techniques – the best 94% of techniques that use a

1-hour update cycle beat all other techniques, and 100% of techniques that use a 1-hour update cycle beat 96% of other techniques (Figure 4-34).

4.5.6 Sensitivity of Ranks to Decisions in Context

The stories sound different when ranks are considered in the context of locking-in certain decision options.

With week-ahead forecasts, in the context of a decision to use a short span, techniques that use the naïve and linear regression algorithms dominate techniques that use other algorithms even more (Figure 4-35). A decision to use a longer span distinguishes techniques that use the naïve algorithm further (Figure 4-36). But, a decision to use the longest span results in no especially dominant techniques (Figure 4-37). Techniques that use support vector regression become increasingly subordinate with decisions to use longer spans.

With day-ahead forecasts, locking in decisions about update cycle reveals more about the effect of the algorithm decision. In the context of a decision to use a short update cycle, techniques that use linear regression do especially well, not so for techniques that use support vector regression (Figure 4-38). Here, the best 40% of techniques that use linear regression beat all others. But, in the context of a decision to use a longer update cycle, the relative performance of techniques switches – techniques that use linear regression do worse and techniques that use support vector regression do better (Figure 4-39). A decision about update cycle dramatically affects the relative performance of techniques that use the naïve algorithm, too – they do worse as the update cycle increases (Figure 4-38, Figure 4-39, Figure 4-40).

4.5.7 Sensitivity of Ranks Based on Multiple Metrics to Decisions

With week-ahead forecasts, techniques ranked by a MAPE/MaxAE/R2 summation rule take on a technique effectiveness signature that shows decisions for longer span tend to be better, and for support vector regression tend to be worse (Figure 4-41).

With day-ahead forecasts, techniques ranked by a MAPE/MaxAE/R2 summation rule show that the decision about clip is less important.

Restricting techniques to only those qualified to be of practical use reveals further variation. Our results assume qualification criteria of MAPE ≤ 5.50%, MaxAE ≤ 12,000 kWh/day, and $R^2 \geq 0.91$. The thresholds are chosen to align with reasonable industry objectives.

With week-ahead forecasts, 905 out of 2,880 techniques satisfy the MAPE criterion, 2,176 satisfy the MaxAE criterion, 59 satisfy the R2 threshold, and 20 of these are fully qualified by satisfying the criteria for all three metrics. Of the fully qualified techniques, MAPE ranged from 2.72% to 4.10%, MaxAE ranged from 2,312 kWh/day to 7,363 kWh/day, and R2 ranged from 0.91 to 0.93. All qualified techniques use a recursive extension rule, 18 use a 1-day update cycle, and 16 use an 18-month span. The top techniques as ranked by MAPE, MaxAE, and R2 within the set of qualified techniques ranked 4th, 6th, and 3rd when ranked by the MAPE/MaxAE/R2 summation rule, respectively (Table 4-3). Qualified techniques do not necessarily perform well across all of MAPE, MaxAE, and R2. Only 2 out of 20 techniques are represented among the best performing 10% of all techniques according to MAPE, MaxAE, and R2, specifically the techniques ranked 4th and 6th according to the MAPE/MaxAE/R2 summation rule.

With day-ahead forecasts, 7 out of 648 techniques satisfy the MAPE criterion, 102 satisfy the MaxAE criterion, 85 satisfy the R2 criterion, and 7 of these are fully qualified by satisfying the criteria for all three metrics. Of the fully qualified techniques, MAPE ranged from 4.90% to 5.46%, MaxAE ranged from 436 kWh/hr (= 10, 464 kWh/day) to 457 kWh/hr (= 10,968 kWh/day), and R2 ranged from 0.96 to 0.97. All qualified techniques use the linear regression algorithm, a direct extension rule, and 1-hour update cycle. The top techniques as ranked by MAPE, MaxAE, and R2 ranked 2nd, 4th, and 3rd when ranked by the MAPE/MaxAE/R2 summation rule, respectively (Table 4-4). Qualified techniques here do perform well across all of MAPE, MaxAE, and R2. Five out of 7 techniques are represented among the best performing

5% of all techniques according to MAPE, MaxAE, and R2, and all 7 techniques are represented among the best performing 10% of all techniques according to MAPE, MaxAE, and R2.

4.5.8 Sensitivity of Ranks Based on Penalty Function to Decisions

With week-ahead and day-ahead forecasts, techniques ranked by penalty function look similar to as when ranked by MAPE (Figure 4-43, Figure 4-44, Table 4-5, Table 4-6).

4.6 Insights

We glean the following insights from our results, with the requisite caveat that they are based on one specific smart electric grid dataset and a practically scoped set of experiments.

4.6.1 Decisions

> Forecasting performance is highly sensitive to the combined effects of forecasting process decisions.

In our analysis, we see that forecasting performance depends on the complete vector of 7 decisions. Specifying options for one, two, or three decisions is not sufficient to determine forecast performance. For example, in the case of week-ahead forecasting, specifying a linear regression algorithm class, but leaving other decisions unspecified, actually specifies 480 unique techniques with MAPE scores ranging from 2.47% to 27.45%.

4.6.2 Update Cycle Decision

> The best forecasting techniques tend to use a short update cycle.

The update cycle decision is very important to technique performance. Techniques that use short update cycles usually forecast better than techniques that use long update cycles. Techniques that use update cycle > 1 day forecast about as well any other, on average.

4.6.3 Algorithm Decision

> Many of the best forecasting techniques use the linear regression algorithm.

Many of the best week-ahead forecasts and all the best day-ahead forecasts come from techniques that use linear regression rather than support vector, multilayer perceptron, naïve, or other algorithms, though some techniques that use those algorithms can do relatively well, too. Some techniques that use the naive algorithm forecast better than many techniques that use other algorithms.

4.6.4 Techniques for Week-Ahead Forecasting

> The best week-ahead forecasting techniques use the naïve algorithm when span is short.

> The best week-ahead forecasting techniques use any algorithm except support vector regression when span is long.

4.6.5 Techniques for Day-Ahead Forecasting

> The best day-ahead forecasting techniques use linear regression, direct extension, and short update cycle; other decisions are not as important to performance.

> The best day-ahead forecasting techniques use the naïve or support vector regression algorithm when the update cycle is long.

4.7 Comparison to Benchmark Studies

We further explore application of our analysis methods and computation platform by evaluating relative forecasting performance of techniques across algorithm decisions in the

context of process-level decisions, and comparing the results to those of another study that evaluated them in a narrower context, but also using the Ireland CER data.

We set up our analysis to correspond to Humeau et al.'s and Wijaya et al.'s experiments on day-ahead electricity demand forecasting, where (in our parlance) a technique using the support vector regression (SVM) algorithm outperforms techniques using linear regression or multilayer perceptron (MLP) algorithms, all other decisions being equal, as measured by all of MAPE, RMSE, and NRMSE. [71,99,117] Algorithms are statically tuned with hyper-parameter values pre-computed to optimize each technique's forecast with respect to the three metrics. Other decisions are locked-in.

Span	lock-in	12,672 hours (~18 months)
Algorithm class	3 options	linear regression, multilayer perceptron, support vector regression
Hyper-parameters	lock-in	For mlp: normalize, decay=0, max iterations=100, absolute tolerance=0.0001, relative tolerance=$1x10^{-8}$ For svm: normalize, eps, kernel=radial, γ=1, cost=1000, tolerance=0.01, ε=0.1, shrink
Tuning	lock-in	static tuning
Extension rule	lock-in	direct
Update cycle	lock-in	1 hour
Holdout	lock-in	32.4%
Clip	lock-in	0
Metric	3 options	MAPE, RMSE, NRMSE

4.7.1 Model Instantiation

In our analysis, we construct and explore performance of 1,536 techniques, scored by 3 metrics, by varying decisions for span, algorithm class, extension rule, update cycle, holdout, clip, and metric.

Span	4 options	3 months, 6 months, 12 months, 18 months
Algorithm class	3 options	linear regression, multilayer perceptron, support vector regression
Hyper-parameters	lock-in	For mlp: normalize, decay=0, max iterations=100, absolute tolerance=0.0001, relative tolerance=$1x10^{-8}$ For svm: normalize, eps, kernel=radial, γ=1, cost=1000, tolerance=0.01, ε=0.1, shrink
Tuning	lock-in	static tuning
Extension rule	2 options	direct, recurse
Update cycle	4 options	1 hour, 1 day, 2 days, 3 days
Holdout	4 options	20%, 30%, 40%, 50%
Clip	4 options	0, 2.5%, 5%, 10%
Metric	3 options	MAPE, RMSE, NRMSE

4.7.2 Results

We reproduced the Humeau et al. experiments in Weka and R, varying the algorithm class decision among linear regression, multilayer perceptron, and support vector regression options, locking in all other decisions, and confirmed that the technique using support vector regression out-performed the others. We then re-ran the experiments, adjusting for daylight savings time, temperature reliability, and other data integrity considerations that could potentially inflate performance. Under these new assumptions, the technique using linear regression performed

best (Table 4-7). This held for both static and dynamic tuning. Further, when freed to vary all decisions, the best performing techniques among the 1,536 we analyzed improved performance across all three algorithms and all three metrics. A technique using linear regression again performed best, and the worst technique using linear regression performed much better than the worst techniques using other algorithms, though on average techniques using support vector regression did perform slightly better as scored by MAPE.

4.7.3 Insights

> Relative forecasting performance of algorithms is highly sensitive to the interaction effects of the algorithm decision and other forecasting process decisions.

Comparing our results to those of another benchmark study that explored the same data, we find that the forecasting technique rank order and absolute level of performance changes depending on the extent to which forecasting process decisions are explicitly identified and included in the analysis.

4.8 Implications for Smart Electric Grid Design and Electricity Policy

With potentially more accurate "grid-enabled" forecasting techniques may come the opportunity for new smart electric grid designs that leverage this capability.

4.8.1 Economic Costs in Terms of Penalty Functions

Forecasting process decisions are related to the economic costs of smart electric grid operation. If operators plan for too much electricity, it will be wasted, and that can be expensive. If they plan for too little electricity, they will have to obtain extra electricity from somewhere else on short notice, and that can be expensive, too. If they plan for way too little electricity, there just will not be enough even from somewhere else, and that means an interruption of electric power, a blackout, could occur.

Benefits derived from more accurate forecasts are justified only to the extent that they exceed economic costs. As a highly simplified example, consider what happens to electricity prices during an unexpected heat wave such as occurred in the PJM Interconnection service area on July 13, 2013. [131] Electricity procured in advance was priced at approximately $20 per MWh for the 4am hour to $240 per MWh for the 2pm hour. On that day, though, electricity demand exceeded forecasts by about 20,000 MWh throughout the day, peaking at just under 157,000 MWh. Electricity procured on 5-minute notice to cover the gap was priced about $100 per MWh higher on average, ranging from $20 per MWh to $500 per MWh at different times of the day. 20,000 MWh shortfall x 24 hr x $100 per MWh price premium = $48 million per day. The economic cost of a blackout could cost many times this amount. [133] This gives us a sense for the magnitude of costs associated with electricity demand forecasting performance.

As a next order approximation, a penalty function can serve as a proxy cost function, assuming forecast asymmetric error is proportional to the combined impact of all cost elements. An even better approximation of economic costs requires a cost function defined in terms of all cost elements or their dependencies. With such a cost function, a distribution of economic costs can be calculated across various potential actual demand levels through Monte Carlo simulation or other methods. In turn, an economic cost estimate or optimal economic cost boundary can be calculated.

Formulation of a cost function is out of scope here, as it requires separate research to uncover the relationships between many cost elements not related to forecasting. However, we do suggest the form that the cost function may take, to expose how forecasting process decisions can influence economic cost, and how they are therefore important considerations for smart electric grid design and electricity policy (Figure 4-1).

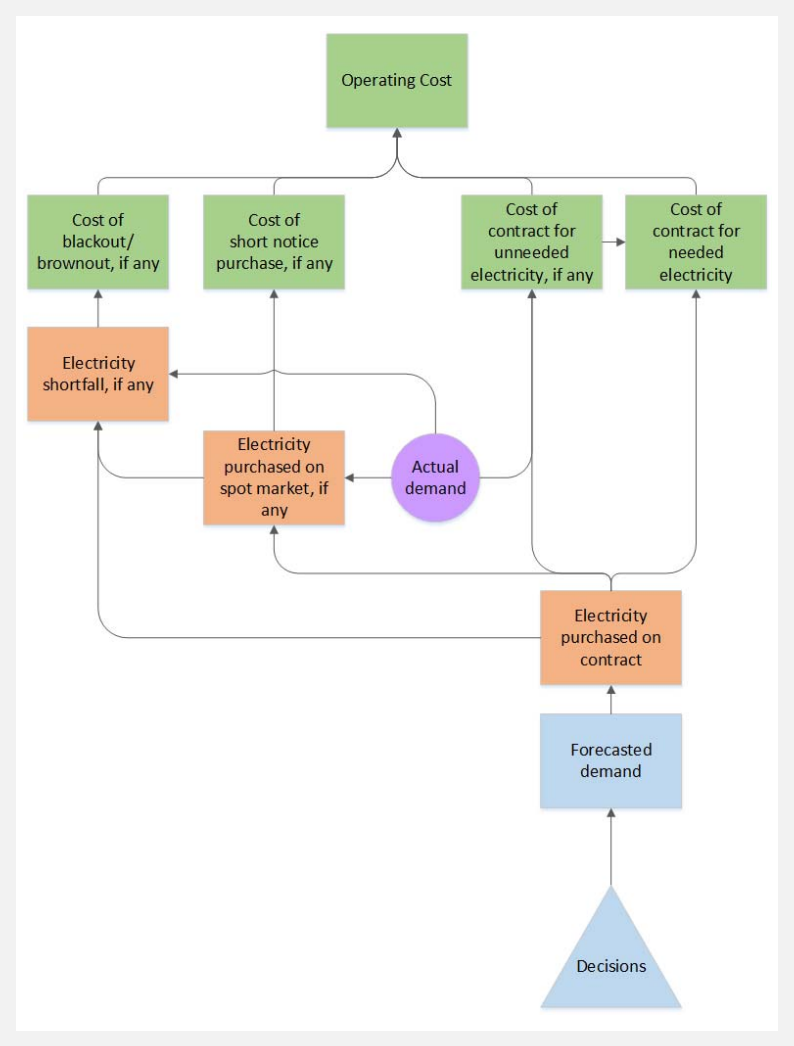


Figure 4-1: Influence diagram of cost function. *Green* indicates a computed economic value. *Orange* indicates other computed values. *Violet* indicates a random event. *Blue* indicates information provided by the forecasting practitioner.

Operating costs is a function of …	• Cost of blackout/brownout, if any • Cost of spot market purchase, if any • Cost of contract for unneeded electricity, if any • Cost of contract for needed electricity
Cost of blackout/brownout, if any, is a function of …	• Electricity shortfall, if any
Cost of short notice purchase, if any, is a function of …	• Electricity purchased on spot market, if any
Cost of contract for unneeded electricity, if any, is a function of …	• Actual demand • Electricity purchased on contract
Cost of contract for needed electricity is a function of …	• Cost of contract for unneeded electricity, if any • Electricity purchased on contract
Electricity shortfall, if any, is a function of …	• Actual demand • Electricity purchased on spot market, if any • Electricity purchased on contract
Electricity purchased on spot market, if any, is a function of …	• Actual demand • Electricity purchased on contract
Actual demand is a random event	
Electricity purchased on contract is a function of …	• Forecasted demand
Forecasted demand is a function of …	• Decisions
Decisions are decisions made by forecasting practitioner	

4.8.2 Economic Costs in Terms of Standard Metrics

If penalty functions do serve as good proxies for cost functions, then it is important that performance scored by MAPE or other standard metrics correlates well with performance scored by penalty functions, so that the large body of research on electricity demand

forecasting using such metrics can be taken as applicable. As noted earlier, MAPE is the de facto metric used to evaluate electricity demand forecasts in many studies.

We see that for the 30 standard metrics and 30 penalty functions we used to evaluate Ireland week-ahead forecasts, standard metric scores across all techniques correlated poorly with those of penalty functions that considered only over-forecasting error, but most correlated extremely well with those of penalty functions that considered either only under-forecasting error or both over- and under-forecasting error (Figure 4-45). For example, MAPE and PT.r10.1.2, the penalty function that assumes a 10% reserve and weighs under-forecasting twice as much as over-forecasting, correlates at r = 0.95.

We also see that the impact of using MAPE instead of a penalty function directly is extremely low, suggesting that MAPE works well as proxy for penalty functions (Table 4-8). The best 20 techniques as scored by MAPE increase PT.r10.1.2 scores by only 0.01, on average, over what techniques scored by PT.r10.1.2 can do.

4.10 Tables and Data Visualizations

Table 4-1: Best performing techniques, ranked by MAPE score, Ireland week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	MAPE.r
168	363	24	naive	direct	24	0.50	0.20	2.33	1
168	534	24	knn	recurse	24	0.30	0.20	2.34	2
168	534	24	naive	direct	24	0.40	0.20	2.47	3
168	534	24	linreg	direct	24	0.30	0.20	2.47	4
168	182	24	naive	recurse	48	0.20	0.20	2.51	5
168	182	24	naive	recurse	72	0.20	0.20	2.51	6
168	534	24	linreg	direct	24	0.40	0.20	2.55	7
168	534	24	naive	direct	24	0.30	0.20	2.57	8
168	534	24	knn	recurse	24	0.30	0.10	2.72	9
168	534	24	tree	recurse	24	0.30	0.20	2.76	10
168	534	24	knn	recurse	48	0.30	0.20	2.78	11
168	534	24	naive	direct	24	0.50	0.20	2.79	12
168	534	24	tree	recurse	24	0.40	0.20	2.82	13
168	534	24	tree	recurse	24	0.40	0.10	2.83	14
168	534	24	knn	recurse	72	0.30	0.20	2.85	15
168	534	24	knn	recurse	24	0.40	0.20	2.87	16
168	534	24	knn	direct	24	0.30	0.20	2.91	17
168	534	24	naive	direct	24	0.40	0.10	2.92	18
168	534	24	knn	direct	48	0.30	0.20	2.94	19
168	534	24	linreg	recurse	24	0.40	0.20	2.94	20

Table 4-2: Best performing techniques, ranked by MAPE score, Ireland day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	MAPE.r
24	534.96	1	linreg	direct	1	0.33	0.20	4.90	1
24	534.96	1	linreg	direct	1	0.33	0.10	5.10	2
24	534.96	1	linreg	direct	1	0.50	0.20	5.12	3
24	363.96	1	linreg	direct	1	0.50	0.20	5.13	4
24	534.96	1	linreg	direct	1	0.50	0.10	5.17	5
24	363.96	1	linreg	direct	1	0.50	0.10	5.32	6
24	534.96	1	linreg	direct	1	0.20	0.20	5.46	7
24	363.96	1	linreg	direct	1	0.33	0.20	5.66	8
24	534.96	1	linreg	direct	1	0.20	0.10	5.71	9
24	534.96	1	linreg	direct	1	0.50	0	5.71	10
24	363.96	1	linreg	direct	1	0.33	0.10	5.79	11
24	182.96	1	linreg	direct	1	0.50	0.20	5.80	12
24	182.96	1	linreg	direct	1	0.33	0.20	5.92	13
24	534.96	1	linreg	direct	1	0.33	0	5.93	14
24	363.96	1	linreg	direct	1	0.50	0	6.09	15
24	182.96	1	linreg	direct	1	0.50	0.10	6.22	16
24	363.96	1	linreg	direct	1	0.20	0.20	6.43	17
24	363.96	1	linreg	direct	1	0.33	0	6.85	18
24	534.96	1	linreg	direct	1	0.20	0	6.88	19
24	182.96	1	linreg	direct	1	0.33	0.10	7.02	20

Table 4-3:

Best performing qualified techniques, ranked by MAPE/MaxAE/R2 scores,
Ireland week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	R2	MaxAE	MAPE.r	R2.r	MaxAE.r	MAPE_MaxAE_R2.r
168	534	24	linreg	recurse	24	0.40	0	3.21	0.93	4202.97	5	2	3	1
168	534	24	linreg	recurse	24	0.40	0.01	3.23	0.93	4202.97	6	3	4	2
168	534	24	linreg	recurse	24	0.50	0	3.39	0.93	4282.50	11	1	5	3
168	534	24	knn	recurse	24	0.30	0.10	2.72	0.91	2695.74	1	15	2	4
168	534	24	linreg	recurse	24	0.50	0.01	3.42	0.93	4282.50	12	4	6	5
168	534	24	tree	recurse	24	0.40	0.10	2.83	0.91	2312.11	2	20	1	6
168	534	24	naive	recurse	24	0.50	0	3.20	0.92	5472.92	4	5	15	7
168	534	24	naive	recurse	24	0.50	0.01	3.19	0.92	5472.92	3	7	16	8
168	534	24	naive	recurse	24	0.40	0	3.27	0.92	5472.92	8	6	13	9
168	534	24	naive	recurse	24	0.40	0.01	3.27	0.92	5472.92	7	8	14	10
168	363	24	naive	recurse	24	0.50	0	3.36	0.91	5472.92	9	11	11	11
168	534	24	mlp	recurse	24	0.40	0.01	3.63	0.91	4422.74	13	12	8	12
168	363	24	naive	recurse	24	0.50	0.01	3.36	0.91	5472.92	10	14	12	13
168	534	24	mlp	recurse	48	0.40	0.01	3.64	0.91	4422.74	14	13	10	14
168	534	24	mlp	recurse	24	0.40	0	3.68	0.91	4422.74	15	17	7	15
168	534	24	mlp	recurse	48	0.40	0	3.68	0.91	4422.74	16	18	9	16
168	363	24	linreg	recurse	24	0.50	0	4.08	0.92	5913.37	19	9	17	17
168	363	24	linreg	recurse	24	0.50	0.01	4.10	0.92	5913.37	20	10	18	18
168	534	24	knn	recurse	24	0.50	0	3.86	0.91	7363.63	17	16	19	19
168	534	24	knn	recurse	24	0.50	0.01	3.87	0.91	7363.63	18	19	20	20

Table 4-4:

Best performing qualified techniques, ranked by MAPE/MaxAE/R2 scores,
Ireland day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	R2	MaxAE	MAPE.r	R2.r	MaxAE.r	MAPE_MaxAE_R2.r
24	534.96	1	linreg	direct	1	0.33	0.10	5.10	0.97	439.70	2	3	4	1
24	534.96	1	linreg	direct	1	0.33	0.20	4.90	0.97	439.70	1	4	5	2
24	534.96	1	linreg	direct	1	0.20	0.20	5.46	0.97	439.24	7	1	3	3
24	534.96	1	linreg	direct	1	0.50	0.10	5.17	0.97	436.26	5	6	1	4
24	534.96	1	linreg	direct	1	0.50	0.20	5.12	0.96	436.26	3	7	2	5
24	363.96	1	linreg	direct	1	0.50	0.10	5.32	0.97	457.14	6	2	6	6
24	363.96	1	linreg	direct	1	0.50	0.20	5.13	0.97	457.14	4	5	7	7

Table 4-5:
Best performing techniques, ranked by PT.r10.1.2 penalty function,
Ireland week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	PT.r10.1.2	PT.r10.1.2.r
168	182	24	naive	recurse	24	0.20	0.20	0	1
168	182	24	naive	recurse	24	0.30	0.20	0	2
168	182	24	naive	recurse	48	0.20	0.20	0	3
168	182	24	naive	recurse	72	0.20	0.20	0	4
168	363	24	naive	direct	24	0.50	0.20	0	5
168	534	24	knn	direct	24	0.30	0.20	0	6
168	534	24	knn	recurse	24	0.30	0.20	0	7
168	534	24	linreg	direct	24	0.40	0.20	0	8
168	534	24	naive	direct	24	0.40	0.20	0	9
168	534	24	naive	direct	24	0.50	0.20	0	10
168	534	24	tree	direct	48	0.40	0.20	0.00	11
168	534	24	tree	direct	72	0.40	0.20	0.00	12
168	534	24	knn	recurse	24	0.40	0.20	0.00	13
168	534	24	knn	recurse	24	0.30	0.10	0.00	14
168	534	24	knn	recurse	24	0.40	0.10	0.00	15
168	534	24	knn	direct	24	0.40	0.20	0.00	16
168	534	24	linreg	recurse	168	0.40	0.20	0.00	17
168	534	24	tree	direct	24	0.40	0.20	0.00	18
168	534	24	tree	direct	120	0.40	0.20	0.00	19
168	534	24	knn	recurse	24	0.20	0.20	0.01	20

Table 4-6:
Best performing techniques, ranked by PT.r10.1.2 penalty function,
Ireland day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	PT.r10.1.2	PT.r10.1.2.r
24	534.96	1	linreg	direct	1	0.50	0.20	0.07	1
24	534.96	1	linreg	direct	1	0.33	0.20	0.07	2
24	534.96	1	linreg	direct	1	0.50	0.10	0.07	3
24	534.96	1	linreg	direct	1	0.33	0.10	0.08	4
24	363.96	1	linreg	direct	1	0.50	0.20	0.08	5
24	534.96	1	linreg	direct	1	0.50	0	0.09	6
24	363.96	1	linreg	direct	1	0.50	0.10	0.09	7
24	534.96	1	linreg	direct	1	0.20	0.20	0.09	8
24	182.96	1	linreg	direct	1	0.33	0.20	0.10	9
24	534.96	1	linreg	direct	1	0.20	0.10	0.10	10
24	182.96	1	linreg	direct	1	0.50	0.20	0.10	11
24	534.96	1	linreg	direct	1	0.33	0	0.10	12
24	363.96	1	linreg	direct	1	0.33	0.20	0.11	13
24	363.96	1	linreg	direct	1	0.33	0.10	0.11	14
24	363.96	1	naive	direct	1	0.20	0.20	0.11	15
24	182.96	1	naive	direct	1	0.33	0.20	0.11	16
24	534.96	1	naive	direct	1	0.33	0.10	0.11	17
24	363.96	1	naive	direct	1	0.50	0.10	0.11	18
24	534.96	1	naive	direct	1	0.33	0.20	0.11	19
24	534.96	1	svm	direct	1	0.33	0.20	0.11	20

Table 4-7: Comparison of forecasting performance across multiple studies.

Techniques	Metric	Algorithm				Algorithm Used by Top-Ranked Technique
		Linear Regression	Multilayer Perceptron	Support Vector Regression static tuning: C=1,000, g=1	Support Vector Regression dynamic tuning: C=10,000, g=1	
Humeau et al. 3 techniques	MAPE	6.1	6.9	4.3		Support Vector Regression
	RMSE	77.0	68.5	54.4		Support Vector Regression
	NRMSE	0.092	0.081	0.064		Support Vector Regression
Humeau et al 3 techniques with adjusted assumptions	MAPE	6.1	8.5	7.9	7.7	Linear Regression
	RMSE	80.0	102.0	101.0	98.1	Linear Regression
	NRMSE	0.094	0.121	0.120	0.115	Linear Regression
Best performing of 2,048 techniques with adjusted assumptions	MAPE	4.8	7.6	7.1		Linear Regression
	RMSE	50.0	79.8	79.6		Linear Regression
	NRMSE	0.068	0.108	0.108		Linear Regression
Worst performing of 2,048 techniques with adjusted assumptions	MAPE	66.3	105.9	102.9		Linear Regression
	RMSE	623.0	791.8	759.0		Linear Regression
	NRMSE	0.650	0.713	0.684		Linear Regression
Mean performance of 2,048 technqiues with adjusted assumptions	MAPE	35.1	36.6	32.7		Support Vector Regression
	RMSE	321.7	338.3	324.9		Linear Regression
	NRMSE	0.336	0.353	0.340		Linear Regression

Table 4-8:

Impact on PT.r10.1.2 when best techniques by MAPE are used, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	MAPE.r	PT.r10.1.2	PT.r10.1.2.r	PT.r10.1.2_impact
168	363	24	naive	direct	24	0.50	0.20	2.33	1	0	5	0
168	534	24	knn	recurse	24	0.30	0.20	2.34	2	0	7	0
168	534	24	naive	direct	24	0.40	0.20	2.47	3	0	9	0
168	534	24	linreg	direct	24	0.30	0.20	2.47	4	0.01	24	0.01
168	182	24	naive	recurse	48	0.20	0.20	2.51	5	0	3	0
168	182	24	naive	recurse	72	0.20	0.20	2.51	6	0	4	0
168	534	24	linreg	direct	24	0.40	0.20	2.55	7	0	8	0
168	534	24	naive	direct	24	0.30	0.20	2.57	8	0.01	25	0.01
168	534	24	knn	recurse	24	0.30	0.10	2.72	9	0.00	14	0.00
168	534	24	tree	recurse	24	0.30	0.20	2.76	10	0.02	114	0.02
168	534	24	knn	recurse	48	0.30	0.20	2.78	11	0.01	23	0.00
168	534	24	naive	direct	24	0.50	0.20	2.79	12	0	10	-0.00
168	534	24	tree	recurse	24	0.40	0.20	2.82	13	0.01	54	0.01
168	534	24	tree	recurse	24	0.40	0.10	2.83	14	0.01	57	0.01
168	534	24	knn	recurse	72	0.30	0.20	2.85	15	0.01	41	0.01
168	534	24	knn	recurse	24	0.40	0.20	2.87	16	0.00	13	-0.00
168	534	24	knn	direct	24	0.30	0.20	2.91	17	0	6	-0.00
168	534	24	naive	direct	24	0.40	0.10	2.92	18	0.02	121	0.02
168	534	24	knn	direct	48	0.30	0.20	2.94	19	0.03	188	0.02
168	534	24	linreg	recurse	24	0.40	0.20	2.94	20	0.01	29	0.00
							mean	2.69	10.50	0.01	37.75	0.01

http://bsnscb.com/dublin-wallpapers/38814219.html

Figure 4-2: Dublin, Ireland.

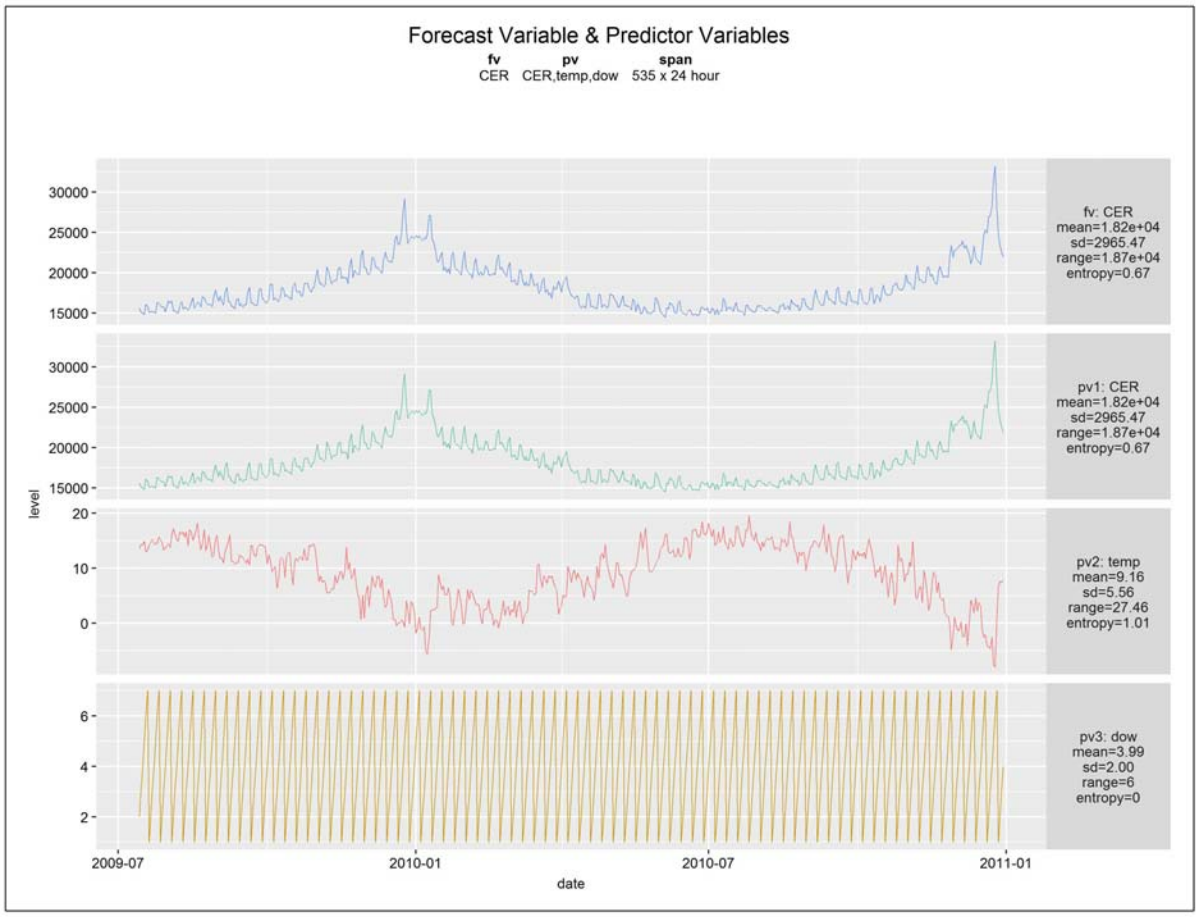


Figure 4-3: Reference and predictor series, Ireland, for week-ahead forecasts. Reference series is aggregated 782-household electricity usage. Predictor series are aggregate electricity usage, temperature, and day of week (integer coded).

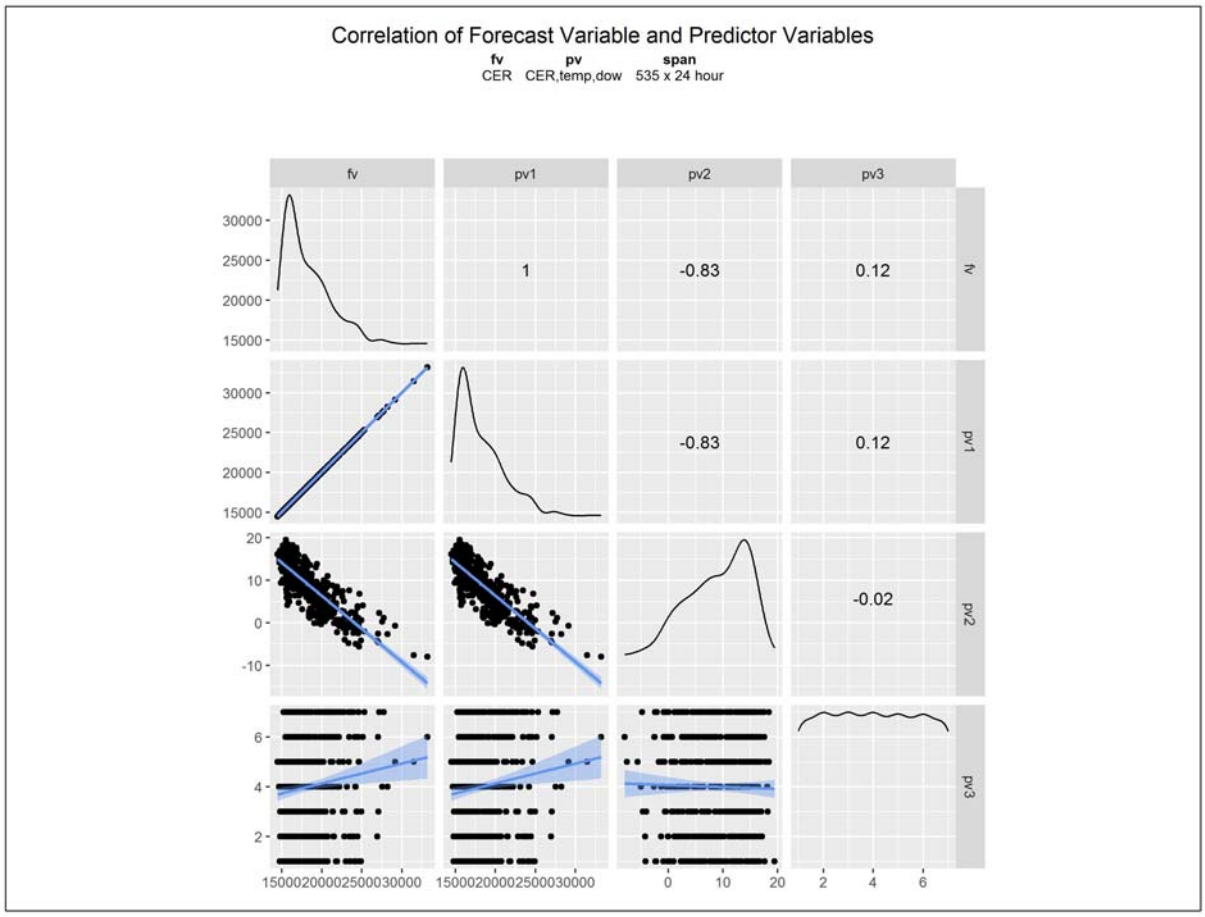


Figure 4-4: Correlations of series, Ireland, for week-ahead forecasts. Reference series is aggregated 782-household electricity usage. Predictor series are aggregate electricity usage, temperature, and day of week (integer coded).

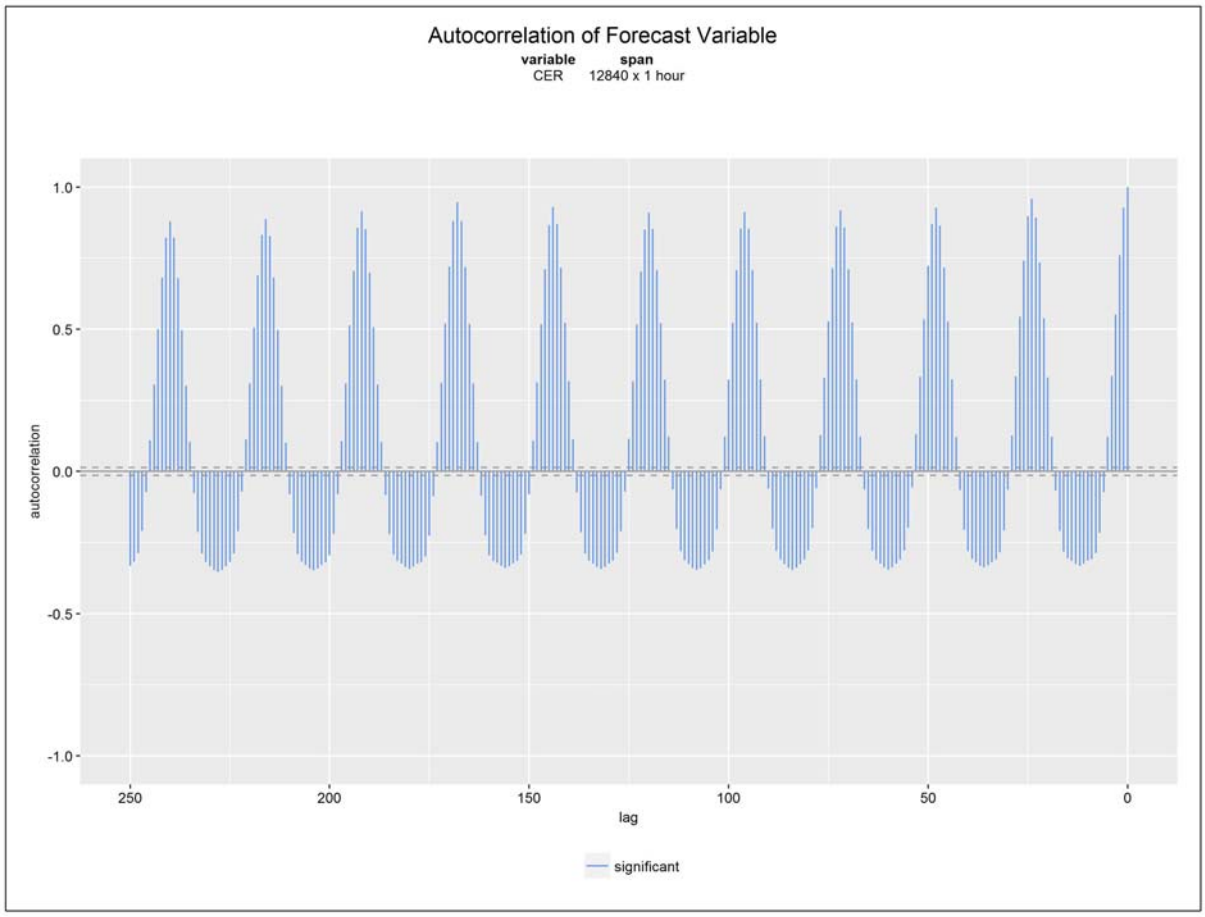


Figure 4-5: Autocorrelation of reference series, Ireland, for day-ahead forecasts. Reference series is aggregate electricity usage.

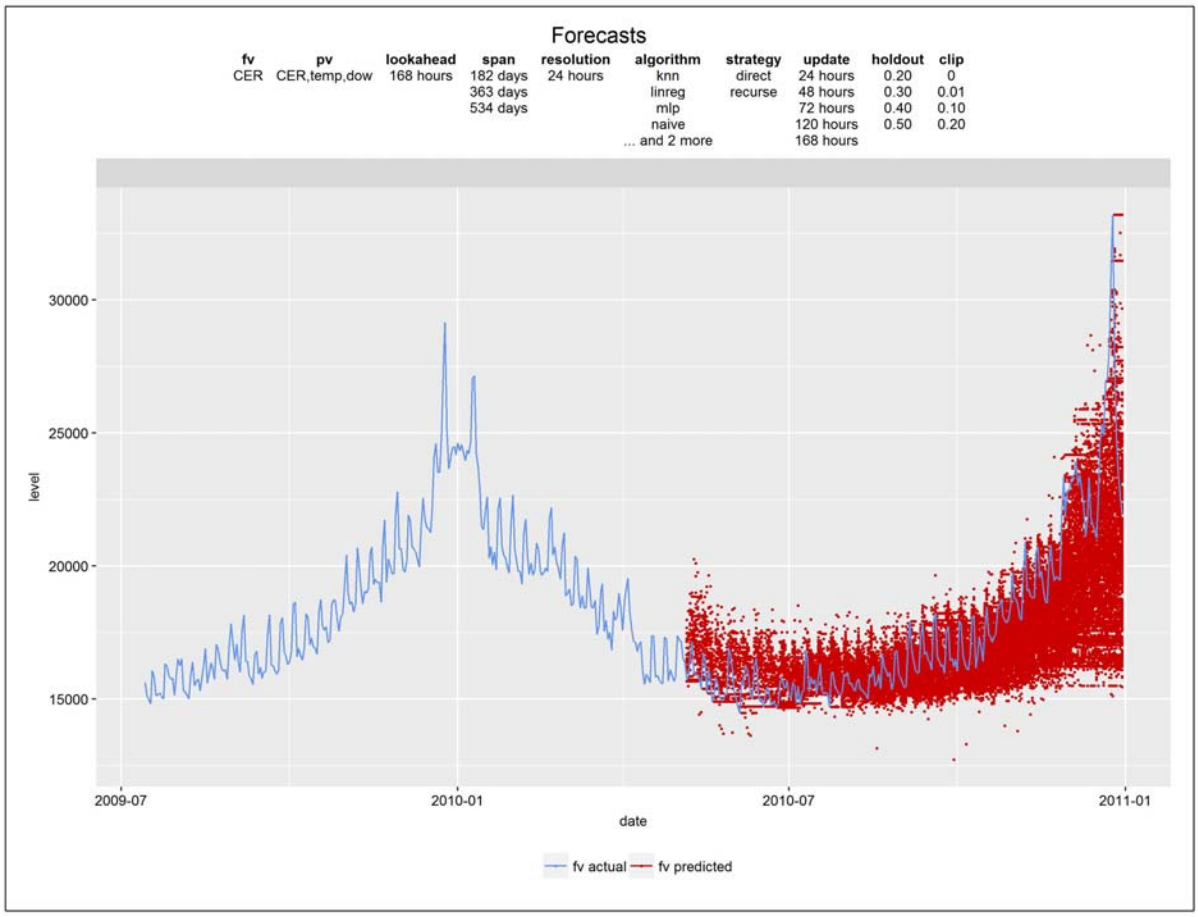


Figure 4-6: Forecasts, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. *Blue* is reference series. *Red* is overlay of forecasts.

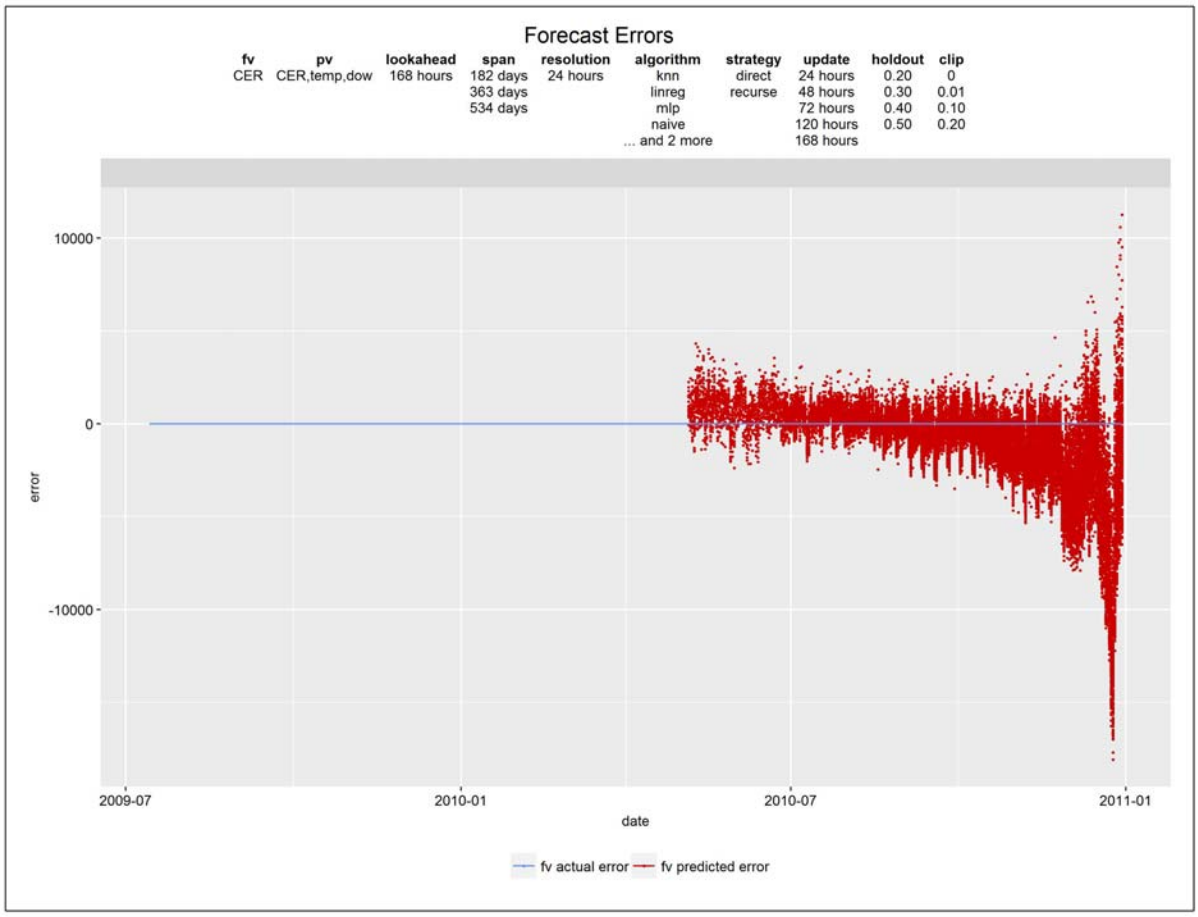


Figure 4-7: Forecast error, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. *Blue* is reference series error (0). *Red* is overlay of forecast errors.

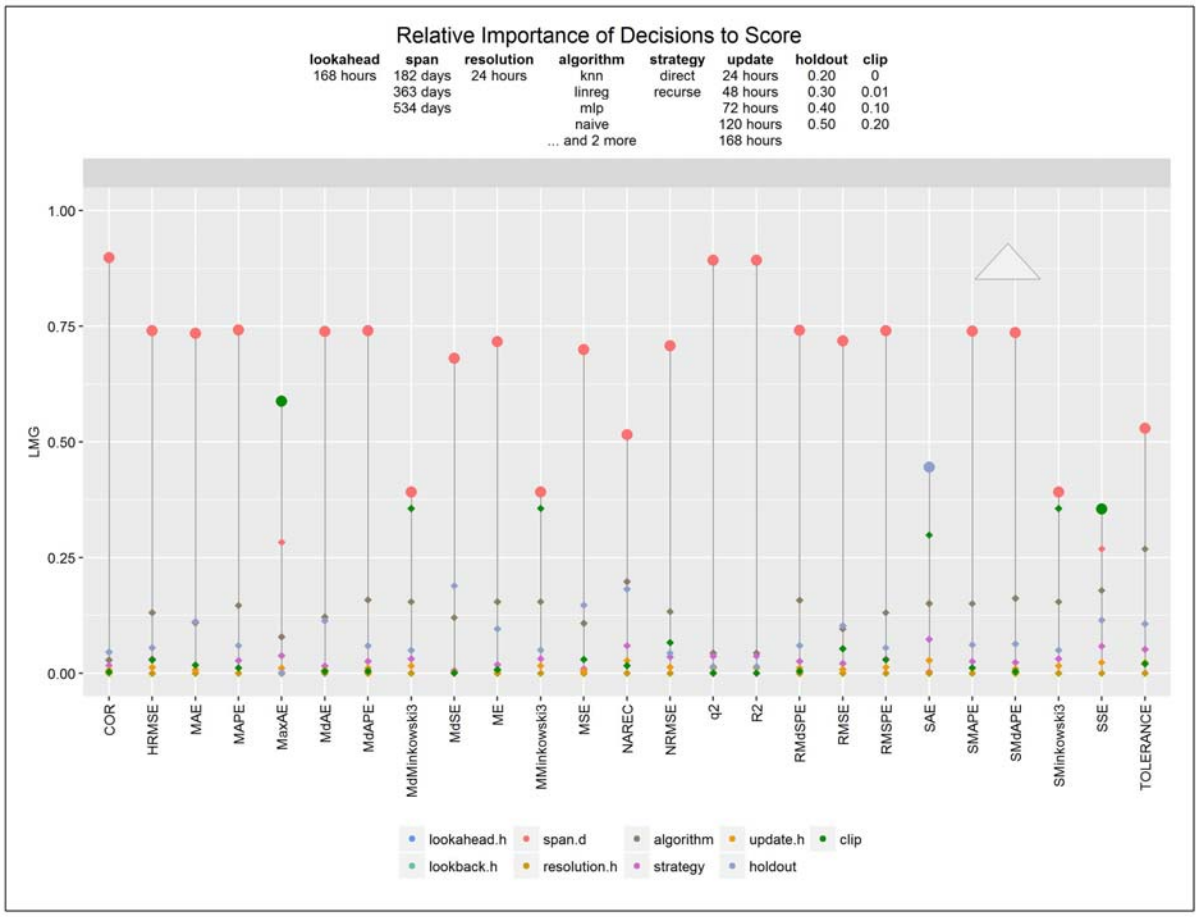


Figure 4-8: Relative importance of decision to metric score variation, across metrics, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. LMG score for each decision is presented along y-axis. Metrics are arranged along x-axis. *Colors* are decision. *Large point* is highest relative importance.

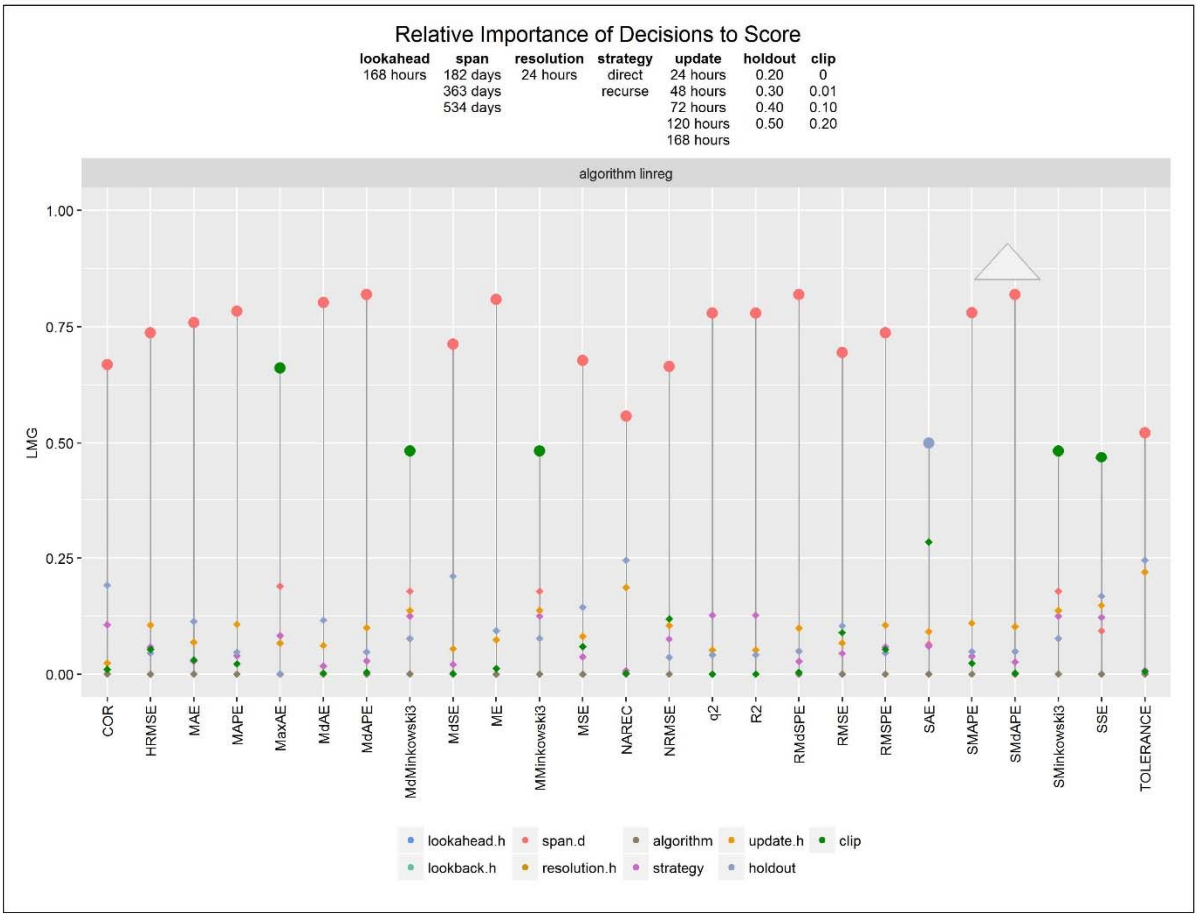


Figure 4-9: Relative importance of decision to metric score variation, linear regression algorithm class only, across metrics, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. LMG score for each decision is presented along y-axis. Metrics are arranged along x-axis. *Colors* are decision. *Large point* is highest relative importance.

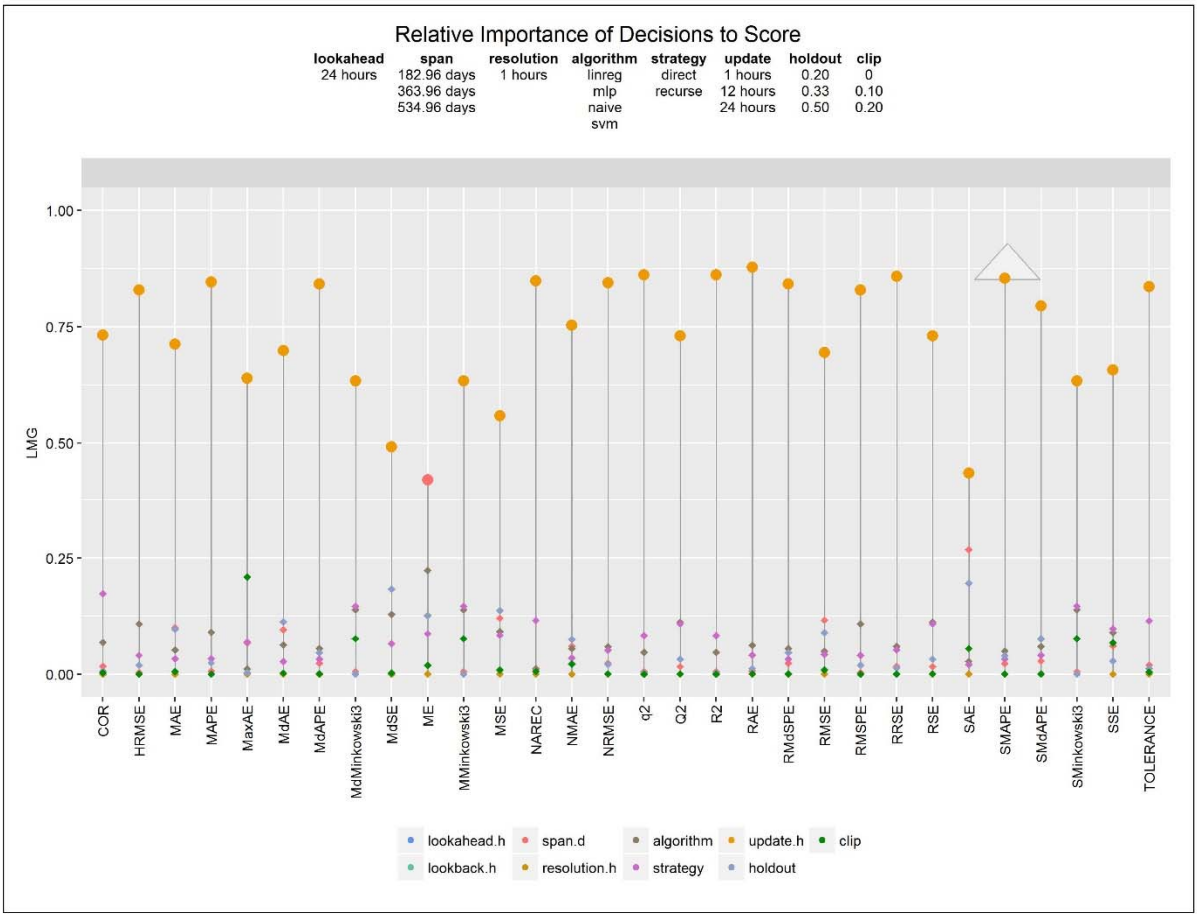


Figure 4-10: Relative importance of decision to metric score variation, across metrics, Ireland, day-ahead forecasts. 648 techniques and forecasts. LMG score for each decision is presented along y-axis. Metrics are arranged along x-axis. *Colors* are decision. *Large point* is highest relative importance.

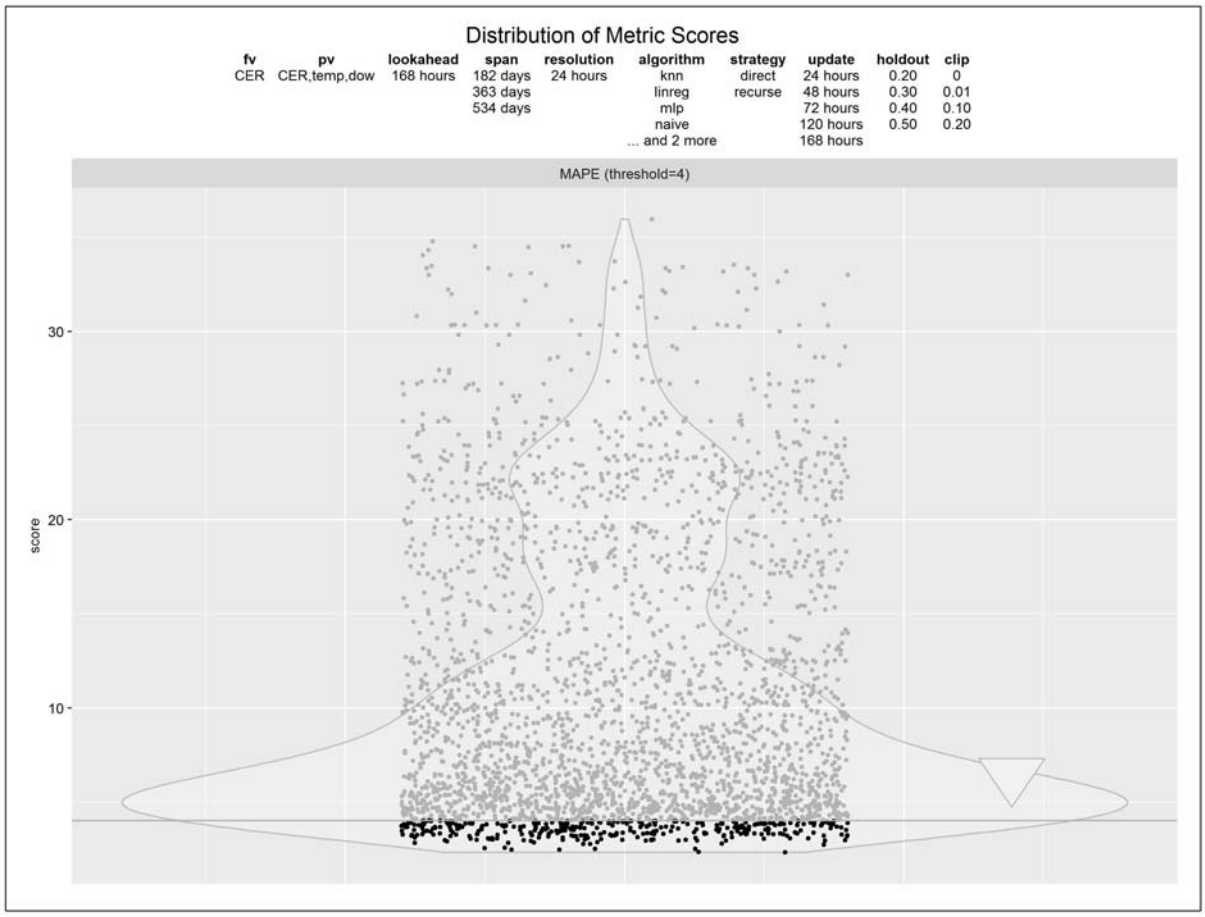


Figure 4-11: Distribution of metric scores, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Metric is MAPE. *Black* is metric score for a qualified technique. *Gray* is metric score for an unqualified technique.

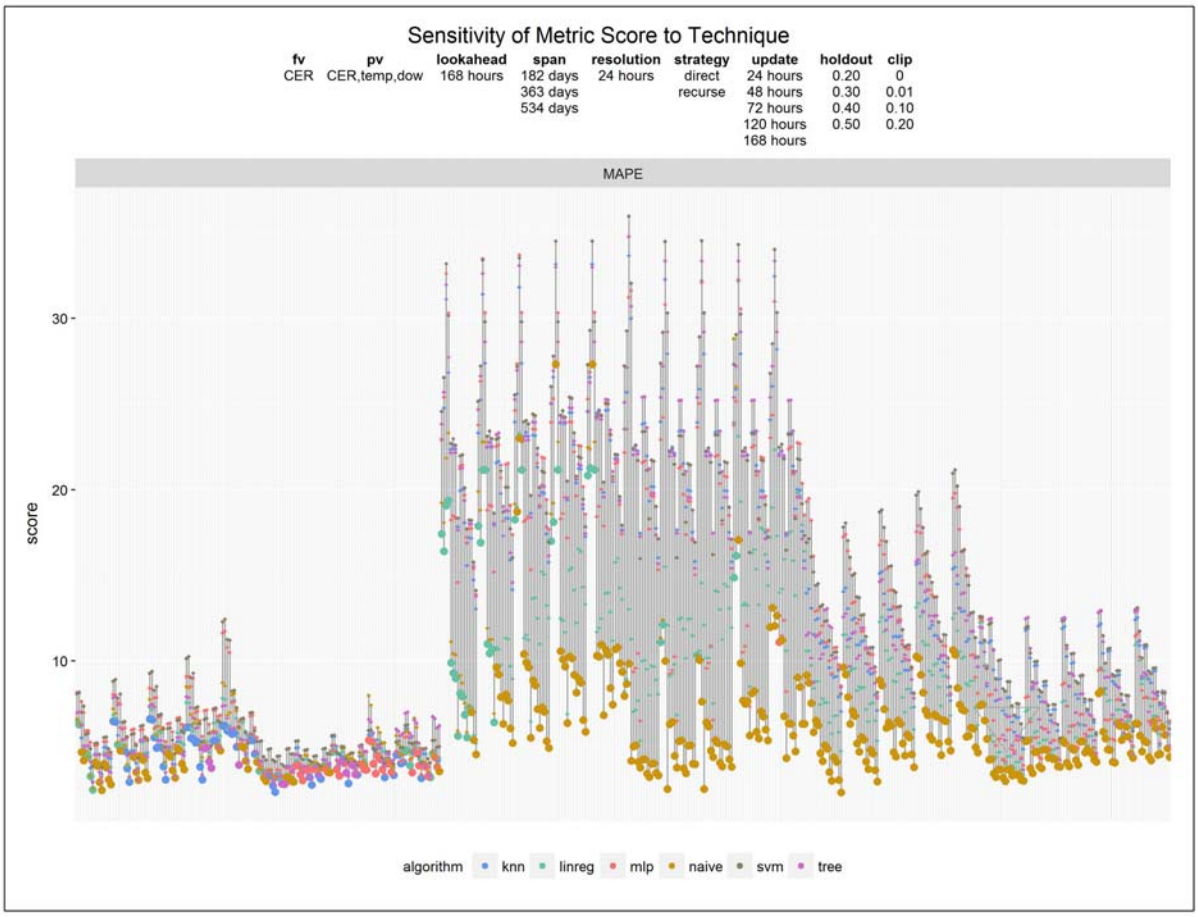


Figure 4-12: Distribution of metric scores across techniques, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each vertical bar represents a family of techniques that differ from each other only by algorithm class. Distances along the vertical bars represent metric scores for specific techniques. Metric is MAPE. *Colors* are algorithm class. *Large point* is metric score of best technique within a family of techniques.

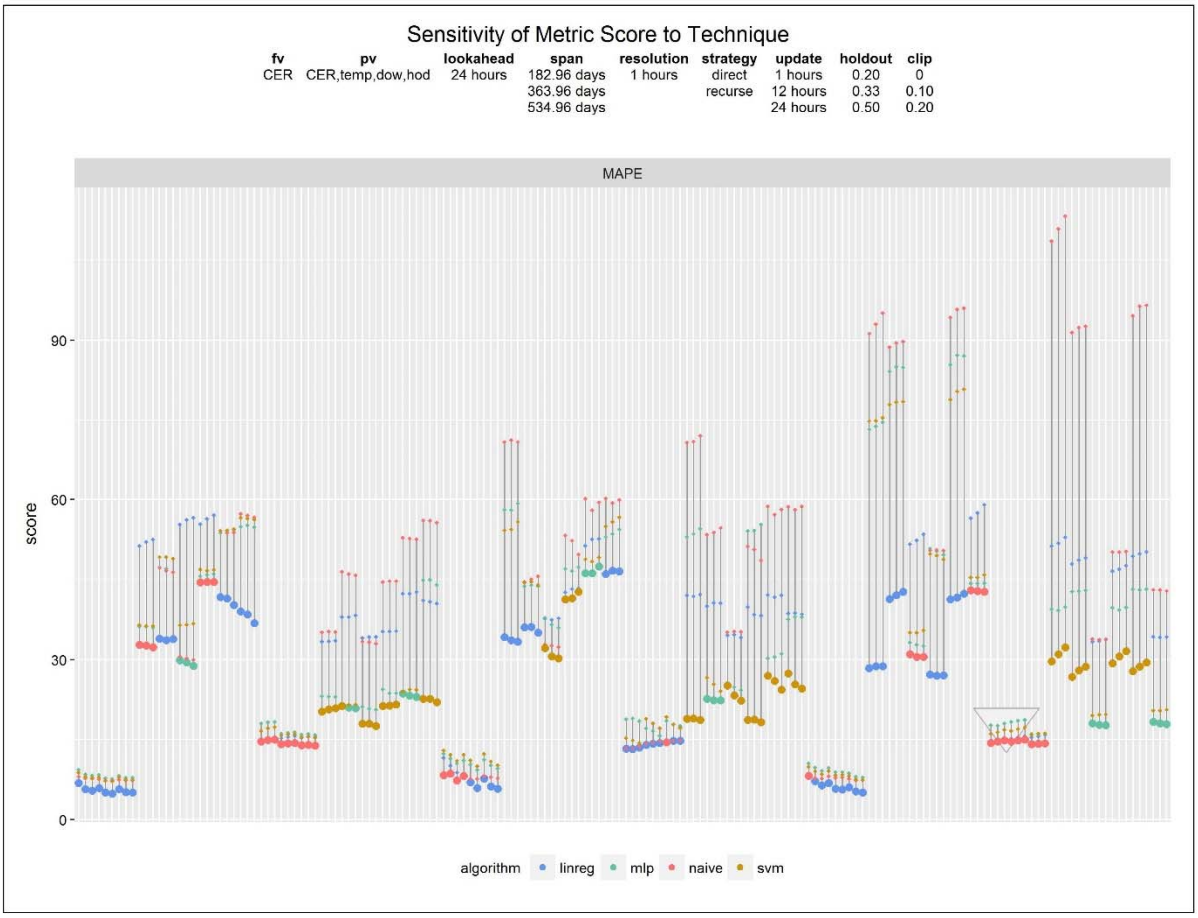


Figure 4-13: Distribution of metric scores across techniques, Ireland, day-ahead forecasts. 648 techniques and forecasts. Each vertical bar represents a family of techniques that differ from each other only by algorithm class. Distances along the vertical bars represent metric scores for specific techniques. Metric is MAPE. *Colors* are algorithm class. *Large point* is metric score of best technique within a family of techniques.

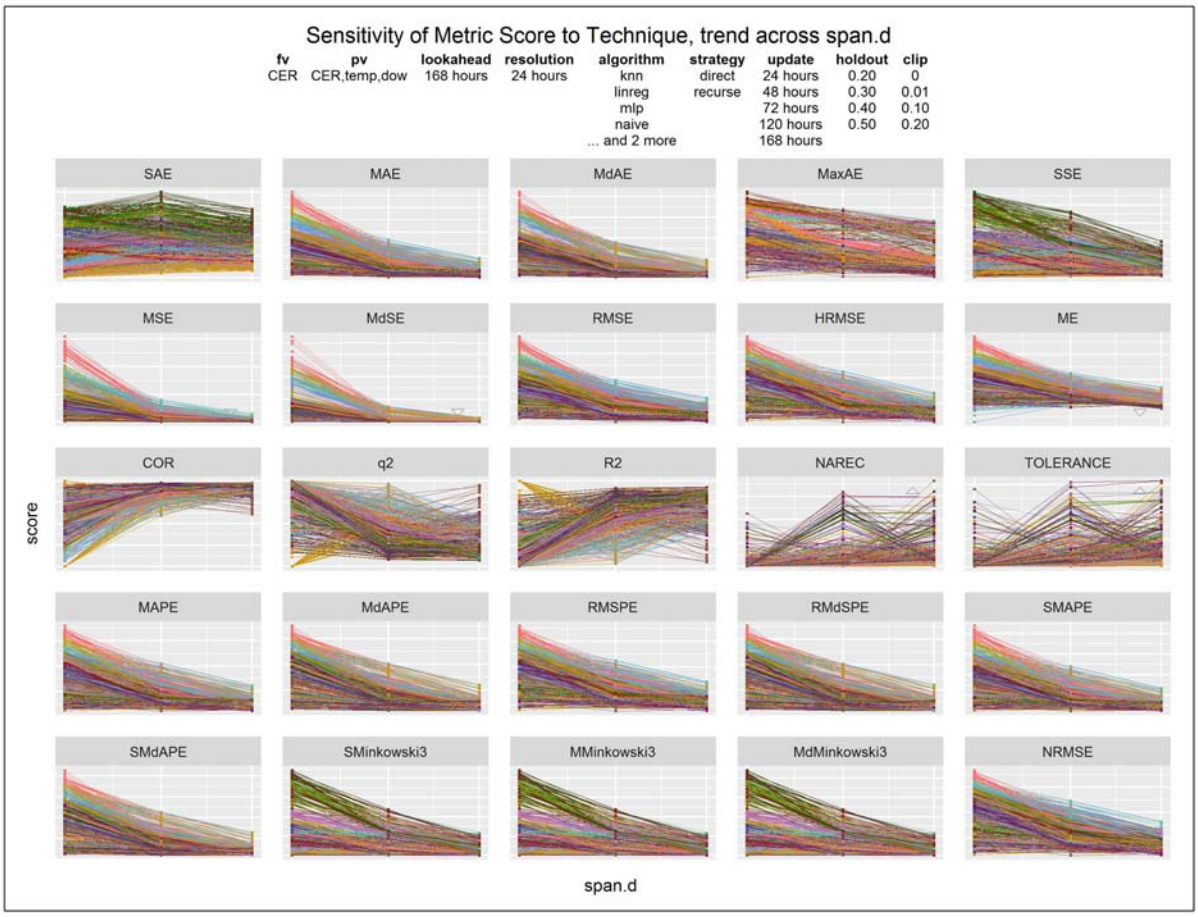


Figure 4-14: Metric score vs. span decision option, split by technique family, for several metrics, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each trend line represents a family of techniques that differ from each other only by span decision option. Spans are arranged sequentially along the x-axis. Metric is as indicated in header strip. *Colors* are technique family.

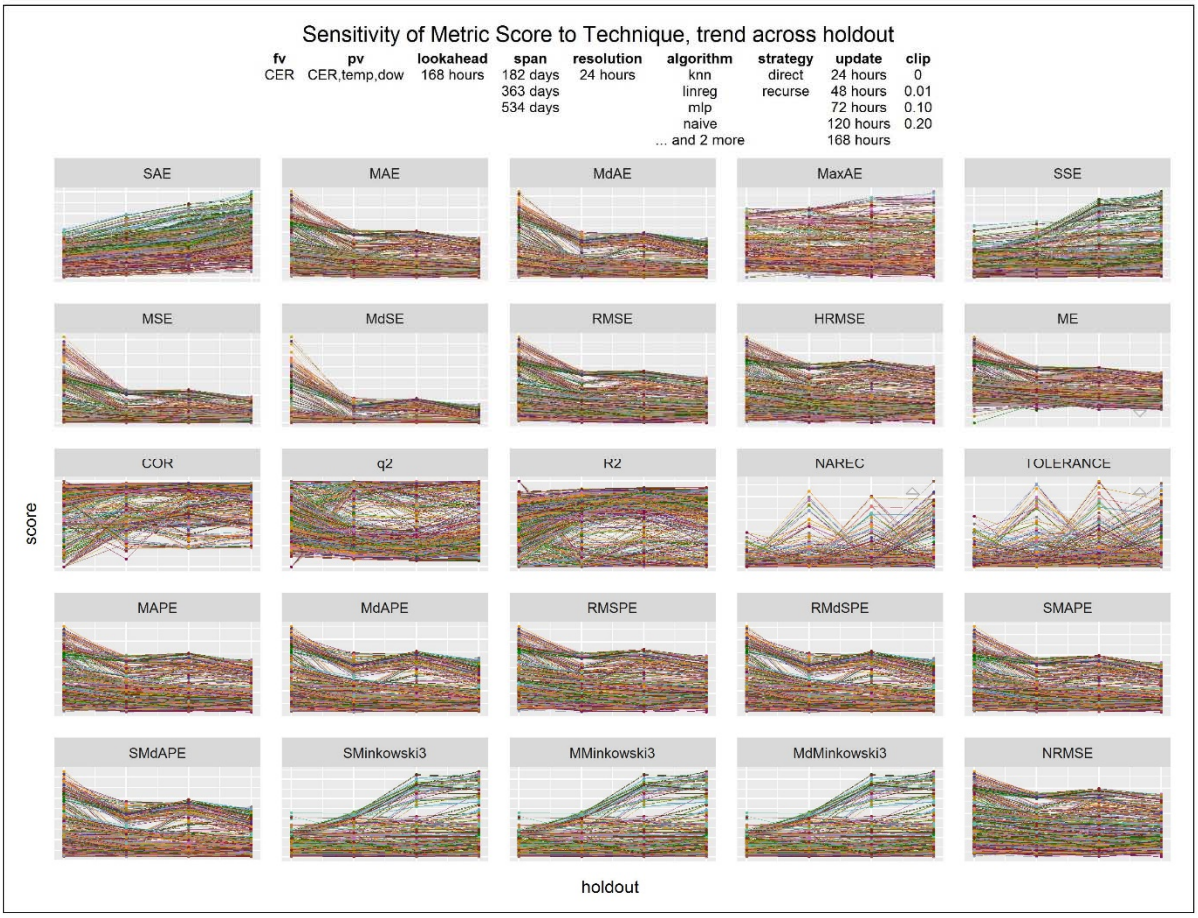


Figure 4-15: Metric score vs. holdout decision option, split by technique family, for several metrics, Ireland, day-ahead forecasts. 2,880 techniques and forecasts. Each trend line represents a family of techniques that differ from each other only by holdout decision option. Holdouts are arranged sequentially along the x-axis. Metric is as indicated in header strip. *Colors* are technique family.

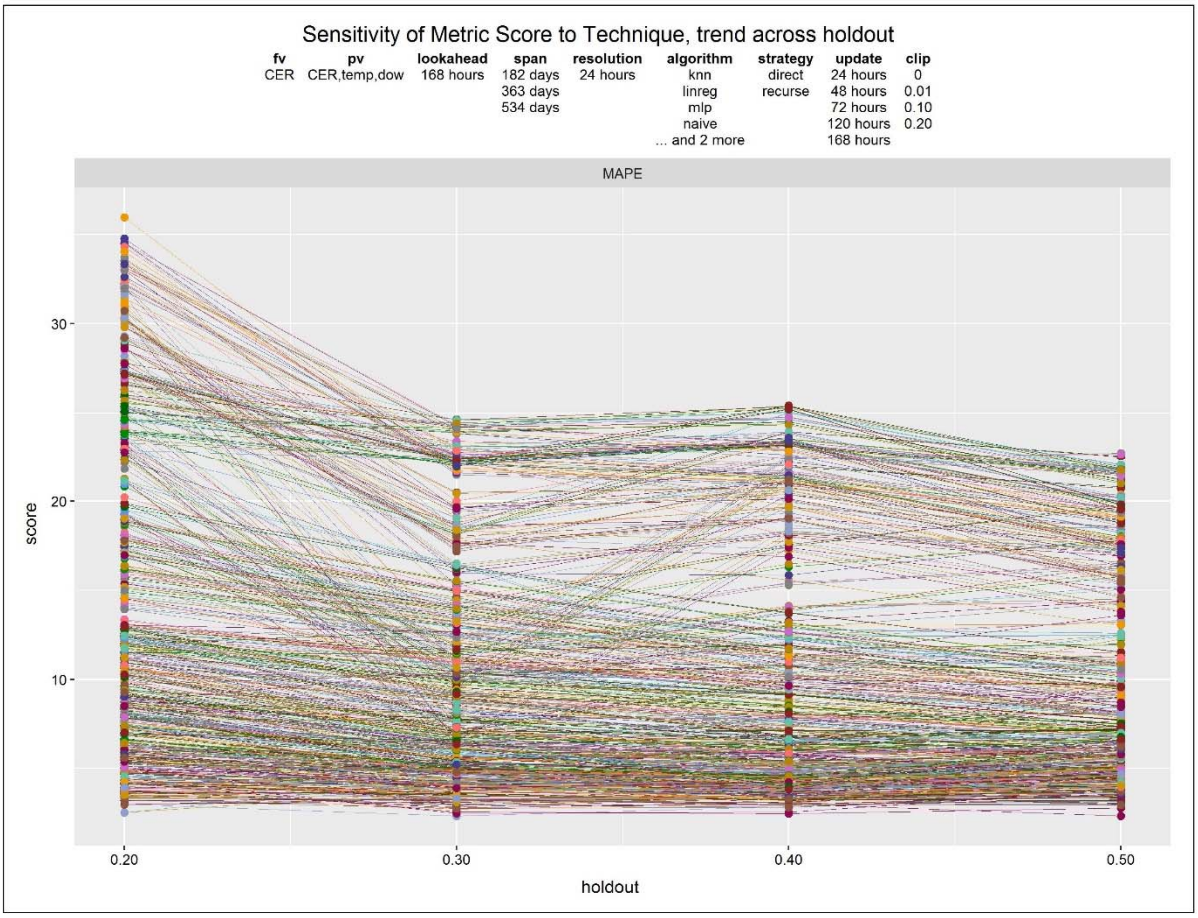


Figure 4-16: Metric score vs. holdout decision option, split by technique family, for several metrics, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each trend line represents a family of techniques that differ from each other only by holdout decision option. Holdouts are arranged sequentially along the x-axis. Metric is MAPE. *Colors* are technique family.

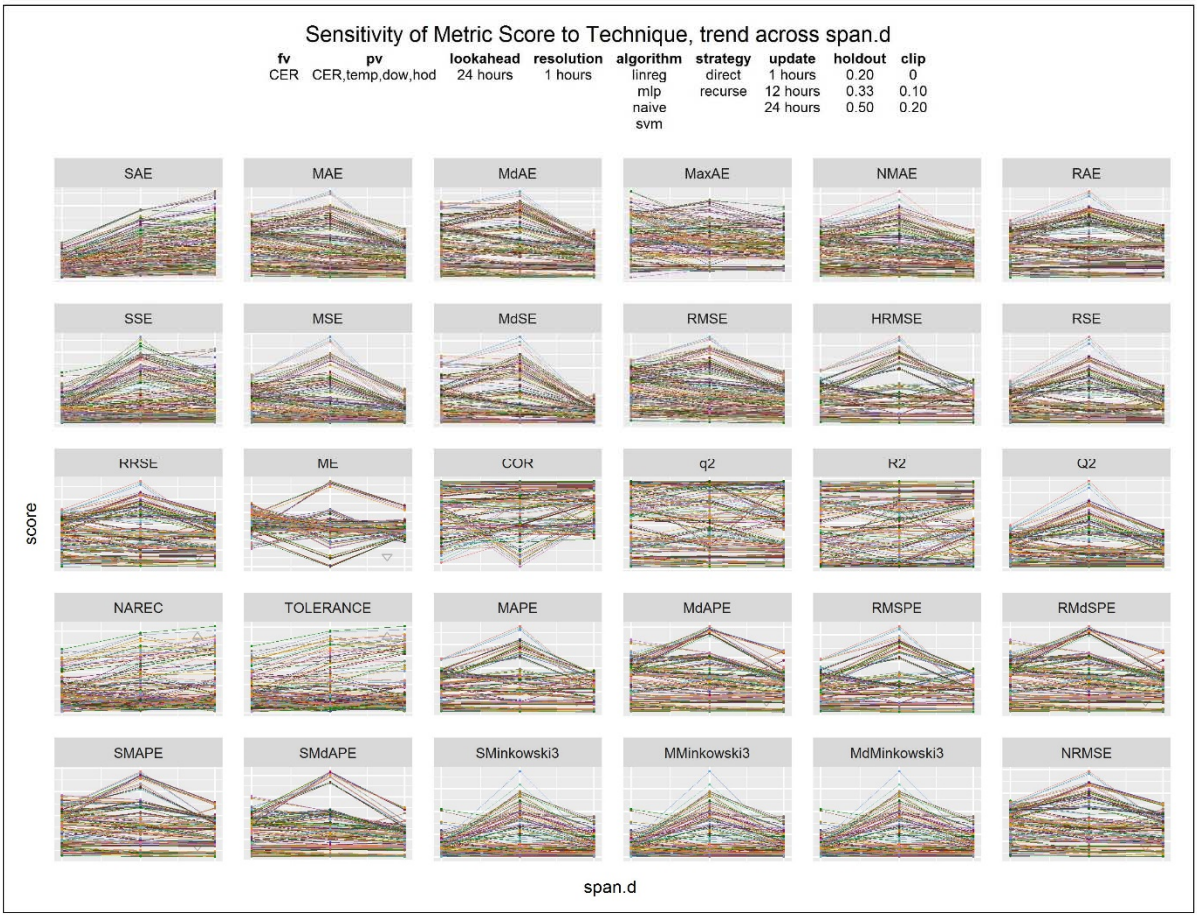


Figure 4-17: Metric score vs. span decision option, split by technique family, for several metrics, Ireland, day-ahead forecasts. 648 techniques and forecasts. Each trend line represents a family of techniques that differ from each other only by span decision option. Spans are arranged sequentially along the x-axis. Metric is as indicated in header strip. *Colors* are technique family.

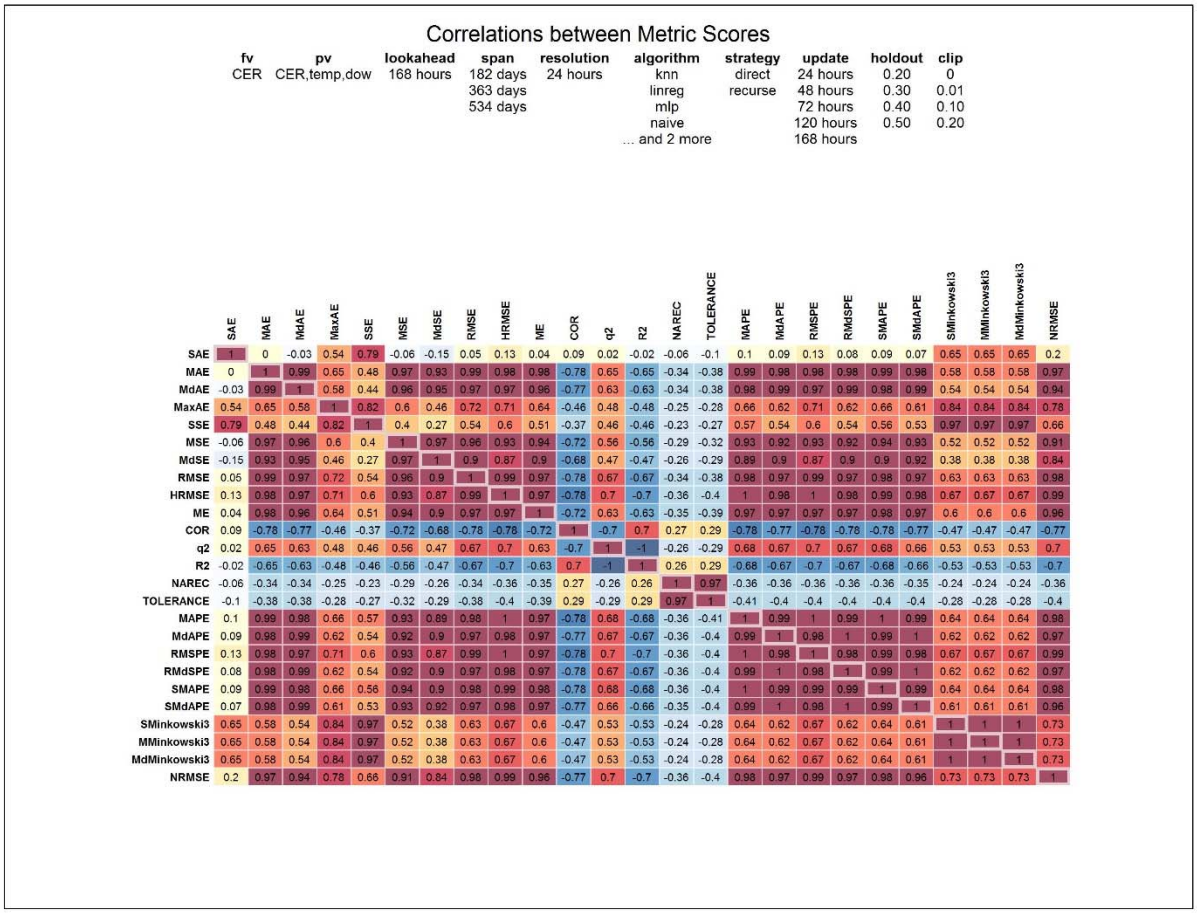


Figure 4-18: Correlations of scores per one metric-to-scores per another metric, across metric pairs, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. For each cell, all techniques are scored per two metrics, and the sequence of scores per the first metric is correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation.

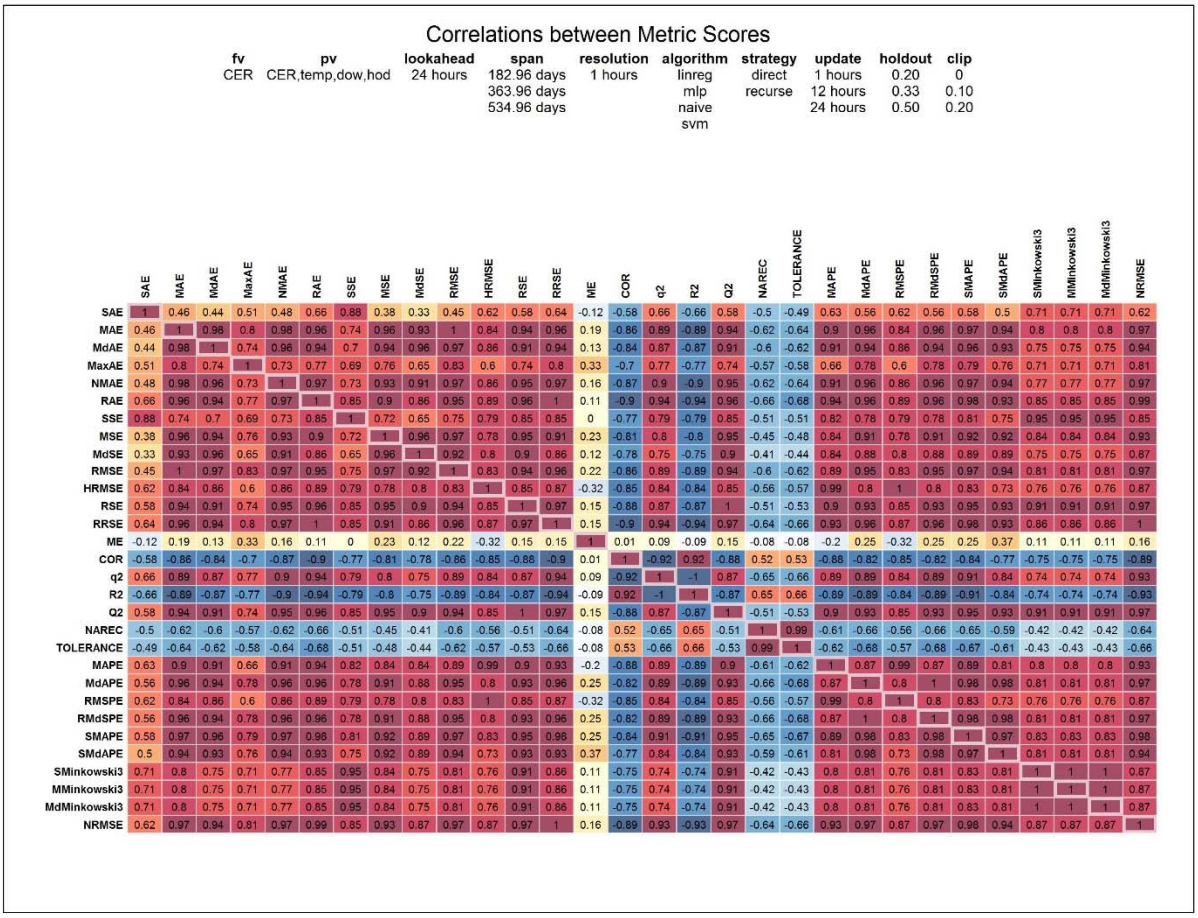


Figure 4-19: Correlations of scores per one metric-to-scores per another metric, across metric pairs, Ireland, day-ahead forecasts. 648 techniques and forecasts. For each cell, all techniques are scored per two metrics, and the sequence of scores per the first metric are correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* indicates strong absolute correlation. *Light* is weak absolute correlation.

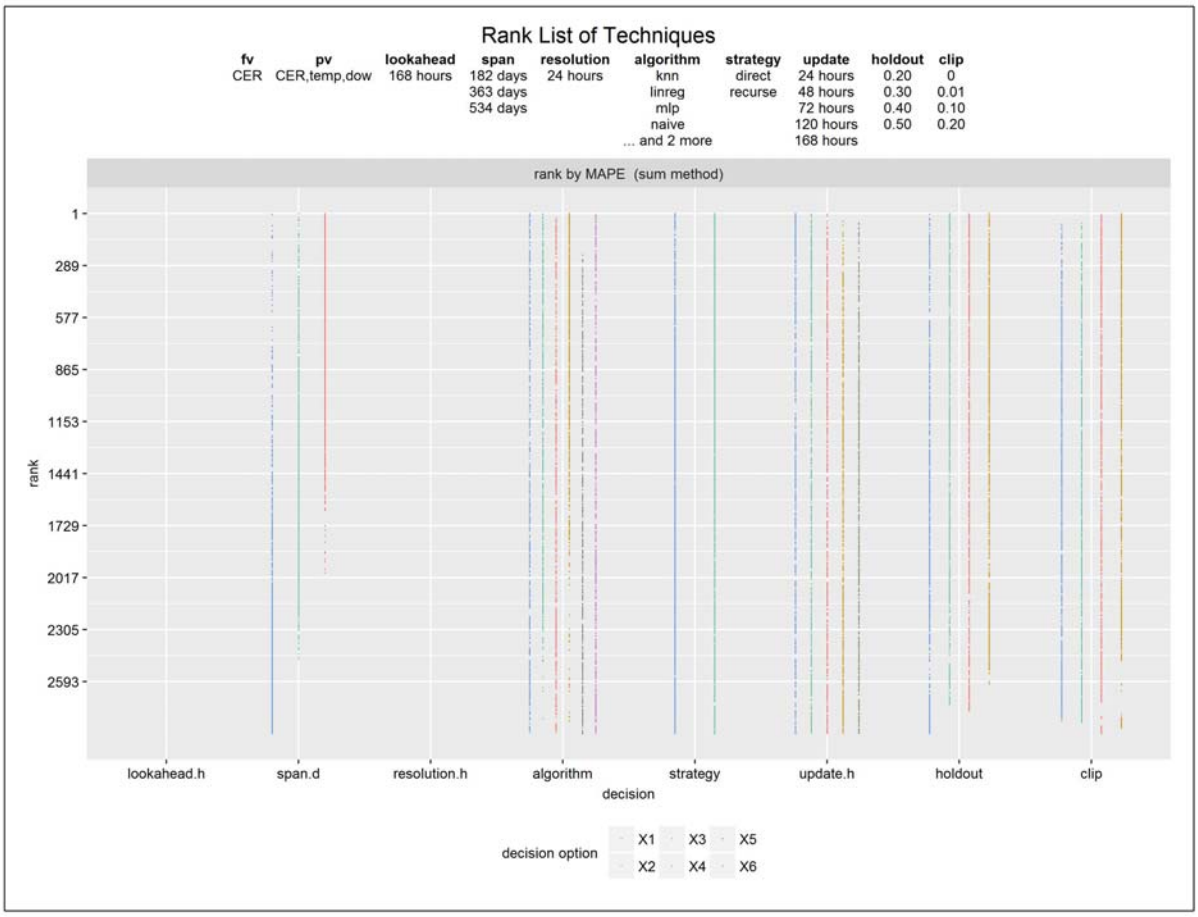


Figure 4-20: List of techniques, rank ordered by performance per MAPE score, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

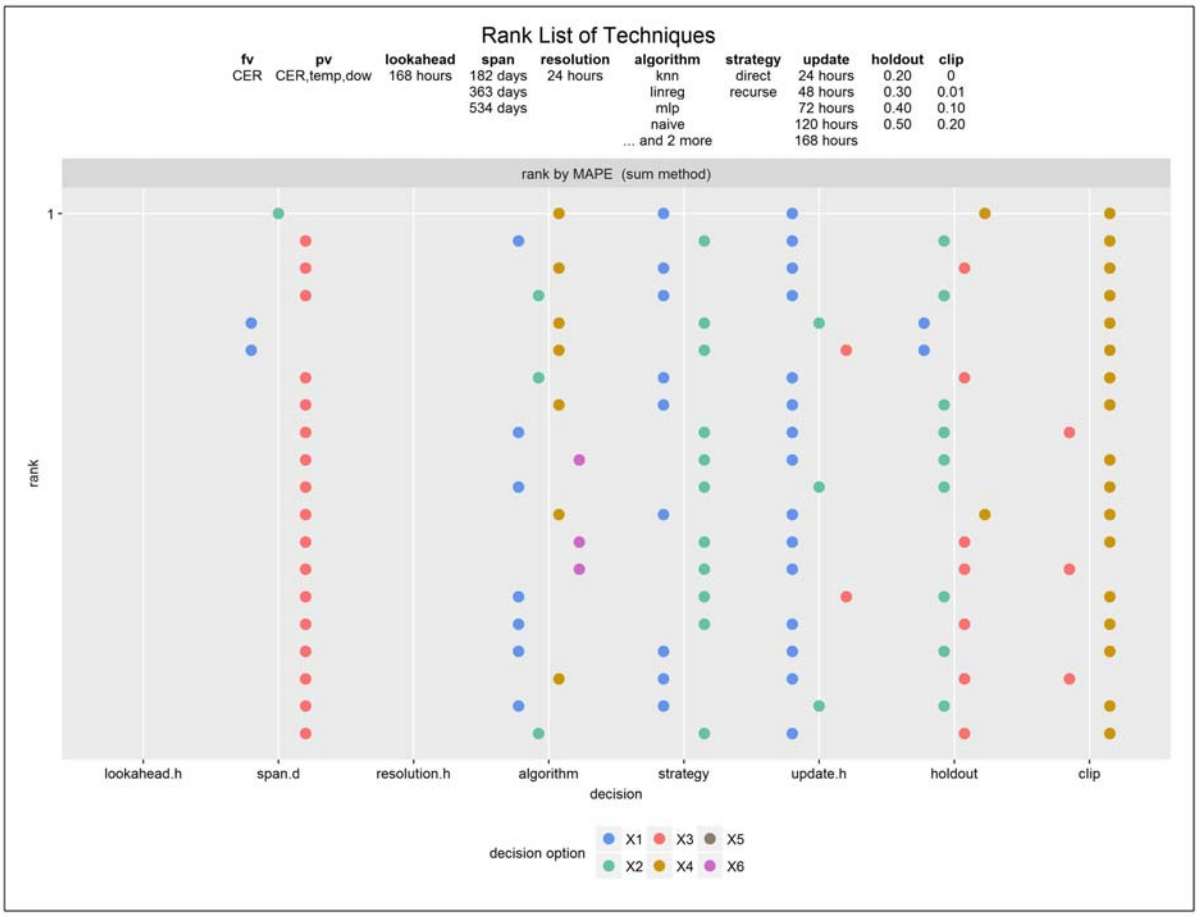


Figure 4-21: List of best techniques, rank ordered by performance per MAPE score, Ireland, week-ahead forecasts. 20 best of 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

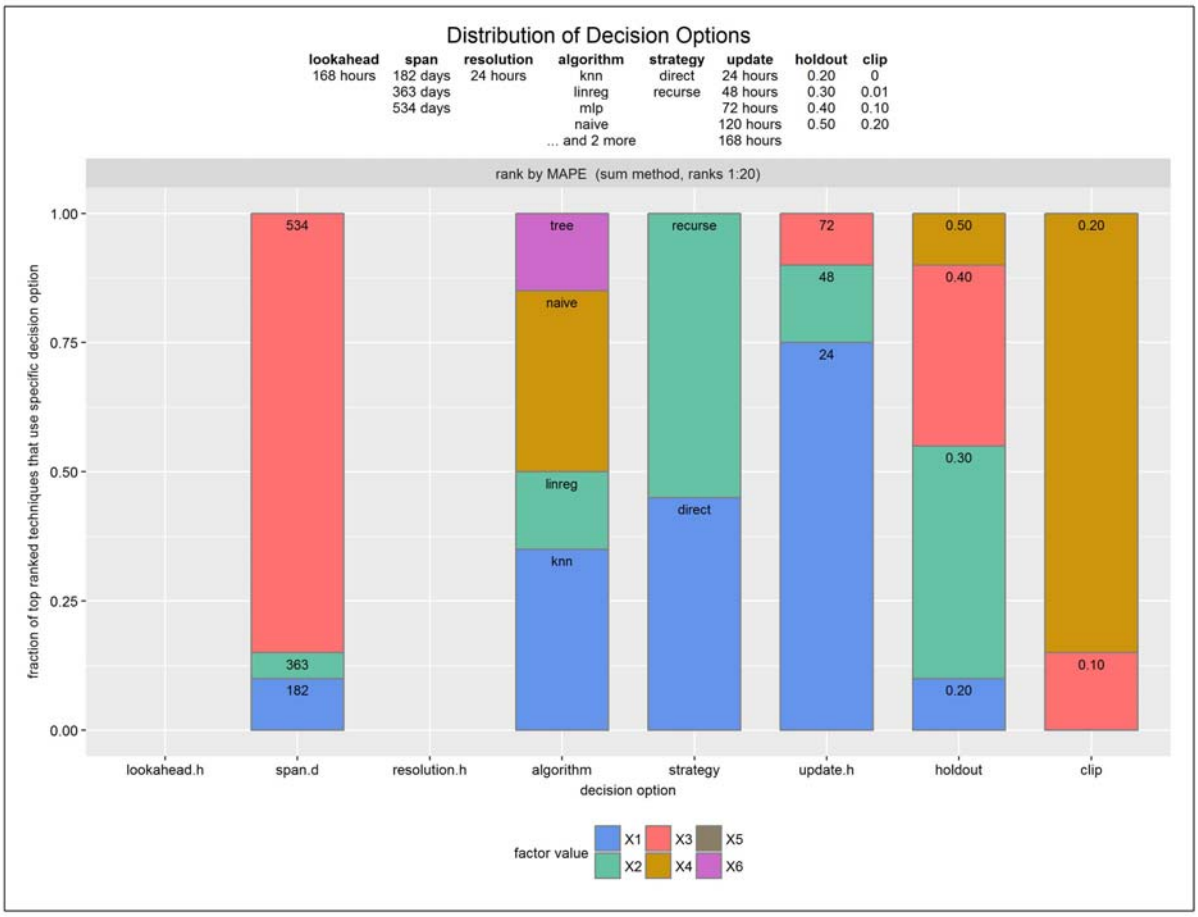


Figure 4-22: Distribution of decision options in best techniques, Ireland, week-ahead forecasts. 20 best of 2,880 techniques and forecasts. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

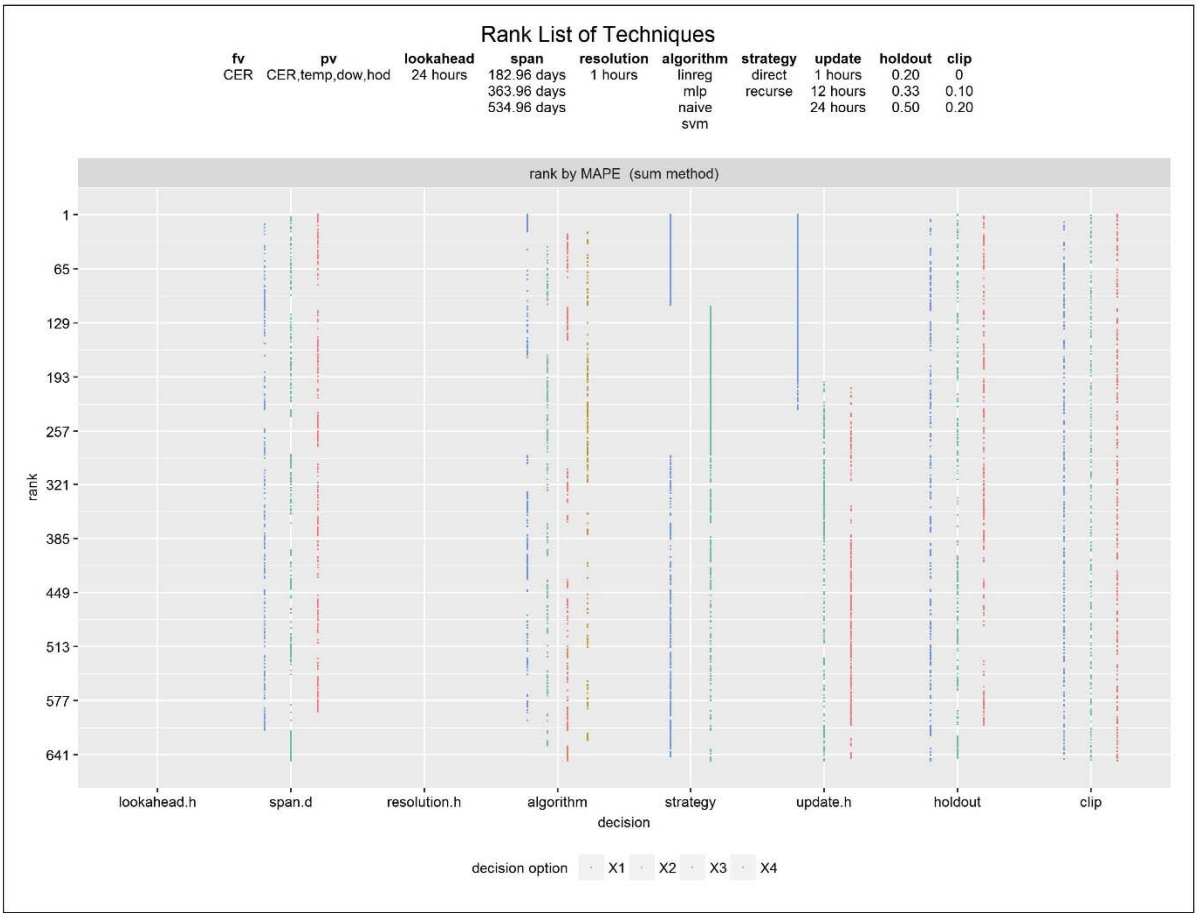


Figure 4-23: List of techniques, rank ordered by performance per MAPE score, Ireland, day-ahead forecasts. 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

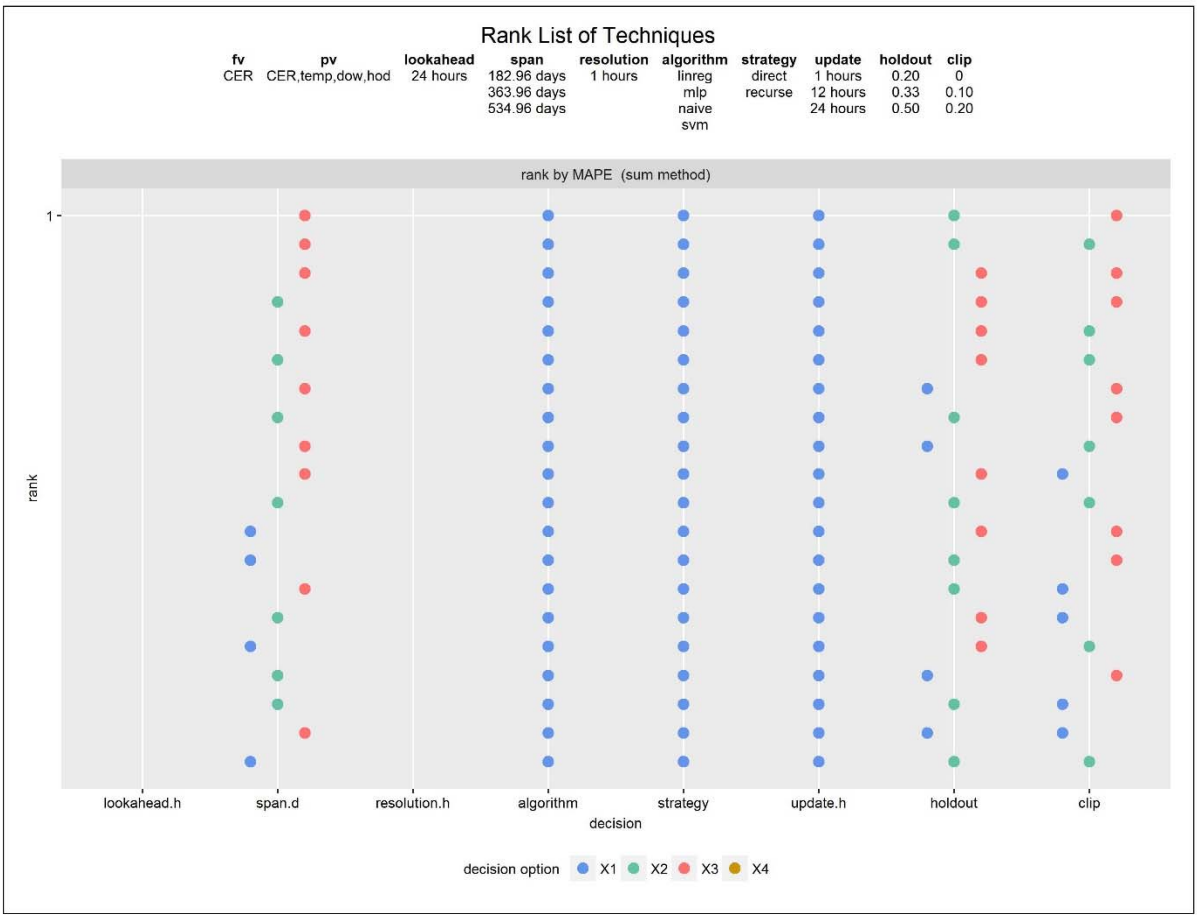


Figure 4-24: List of best techniques, rank ordered by performance per MAPE score, Ireland, day-ahead forecasts. 20 best of 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

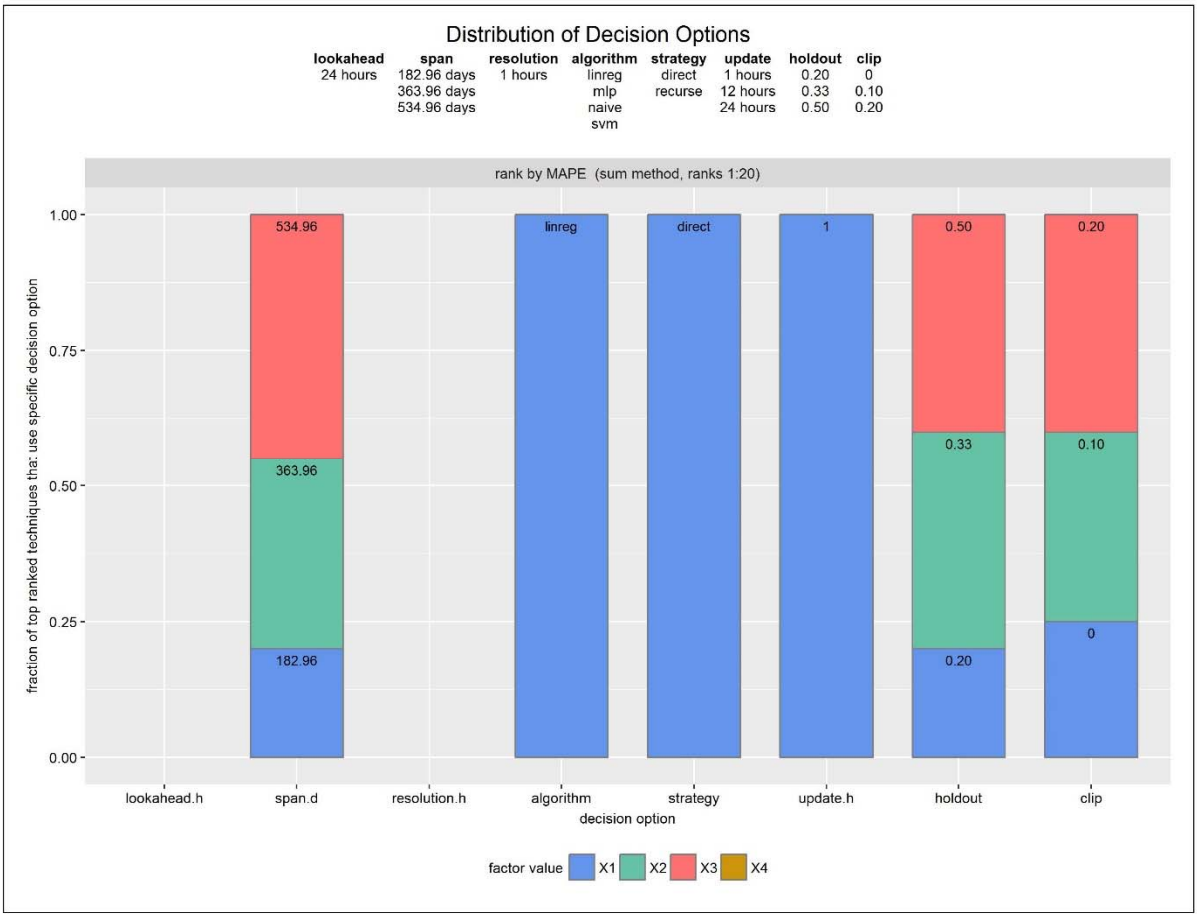


Figure 4-25: Distribution of decision options in best techniques, Ireland, day-ahead forecasts. 20 best of 648 techniques and forecasts. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

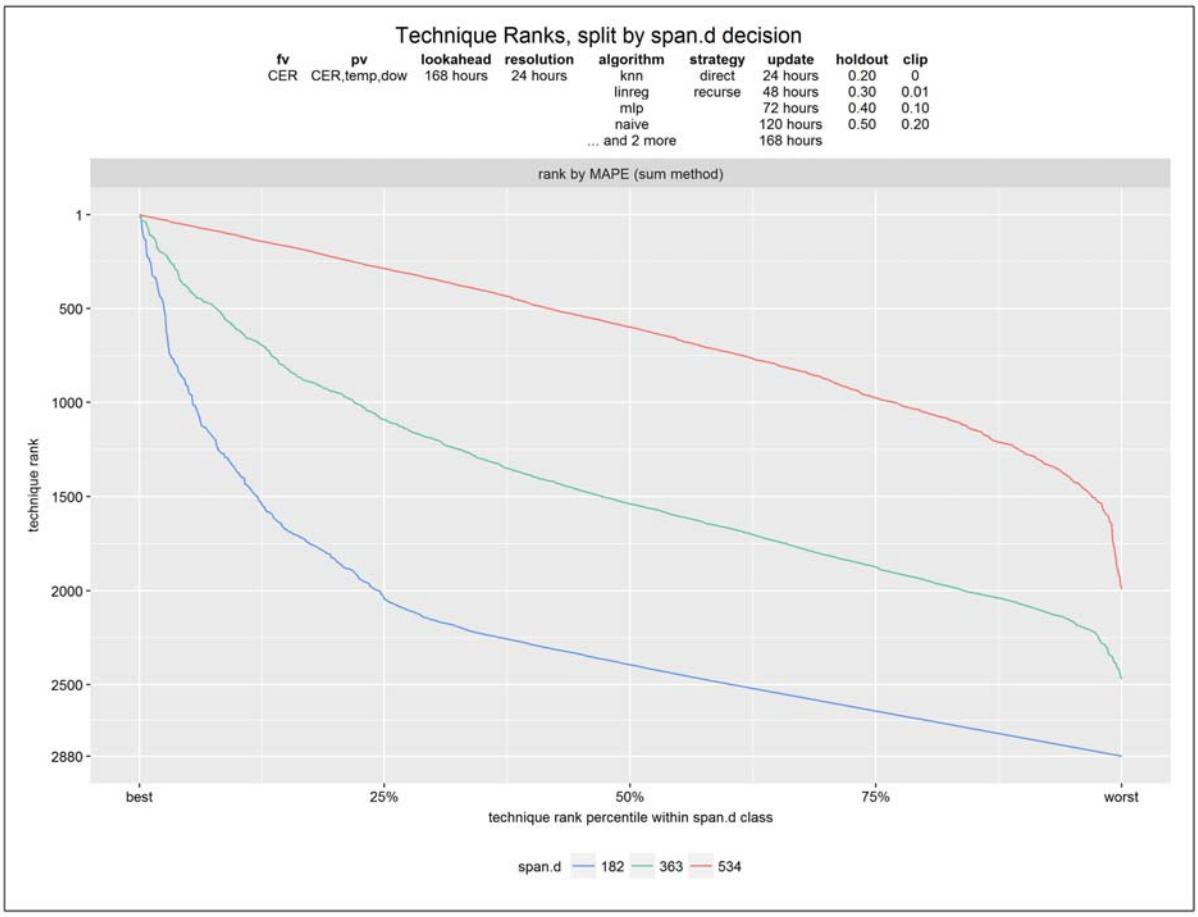


Figure 4-26: Technique rank trend, split by span, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the span decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the span decision option.

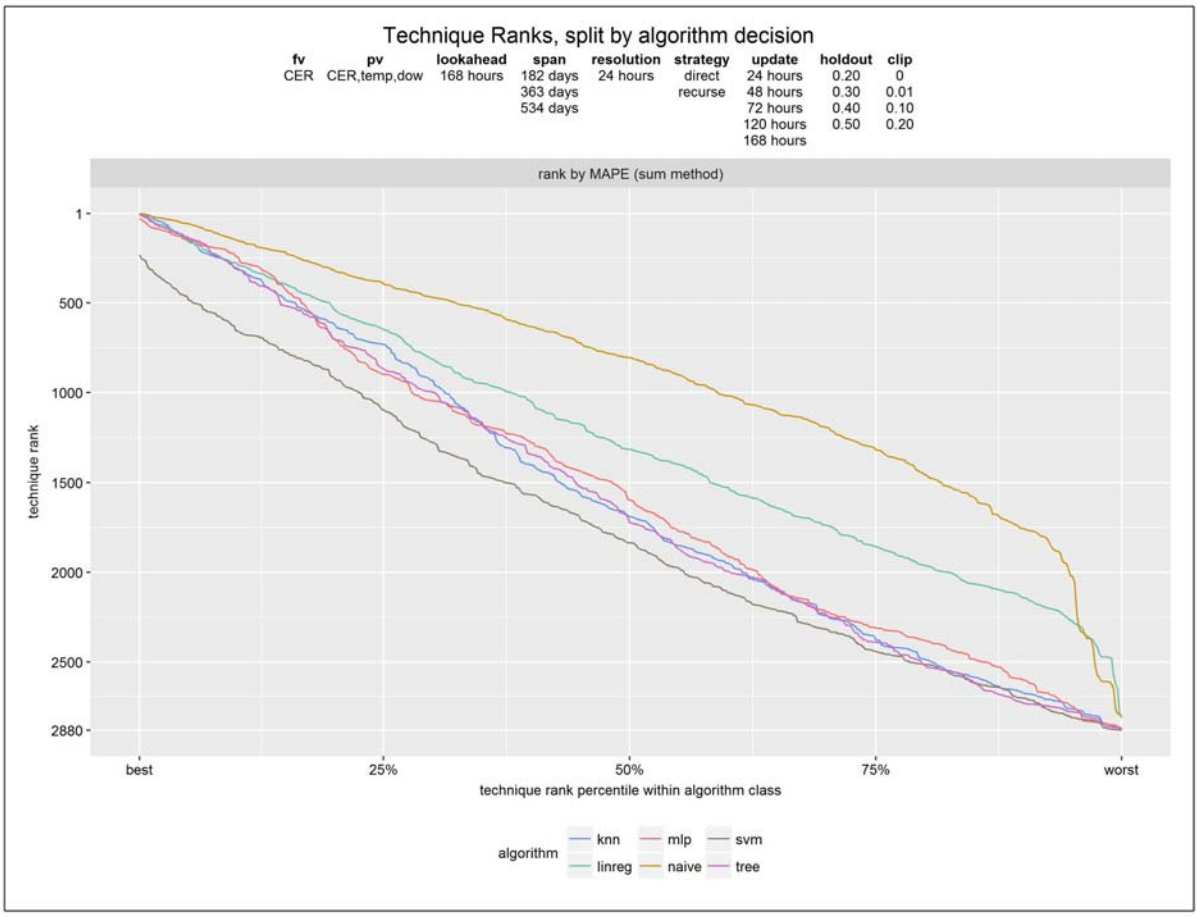


Figure 4-27: Technique rank trend, split by algorithm class, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

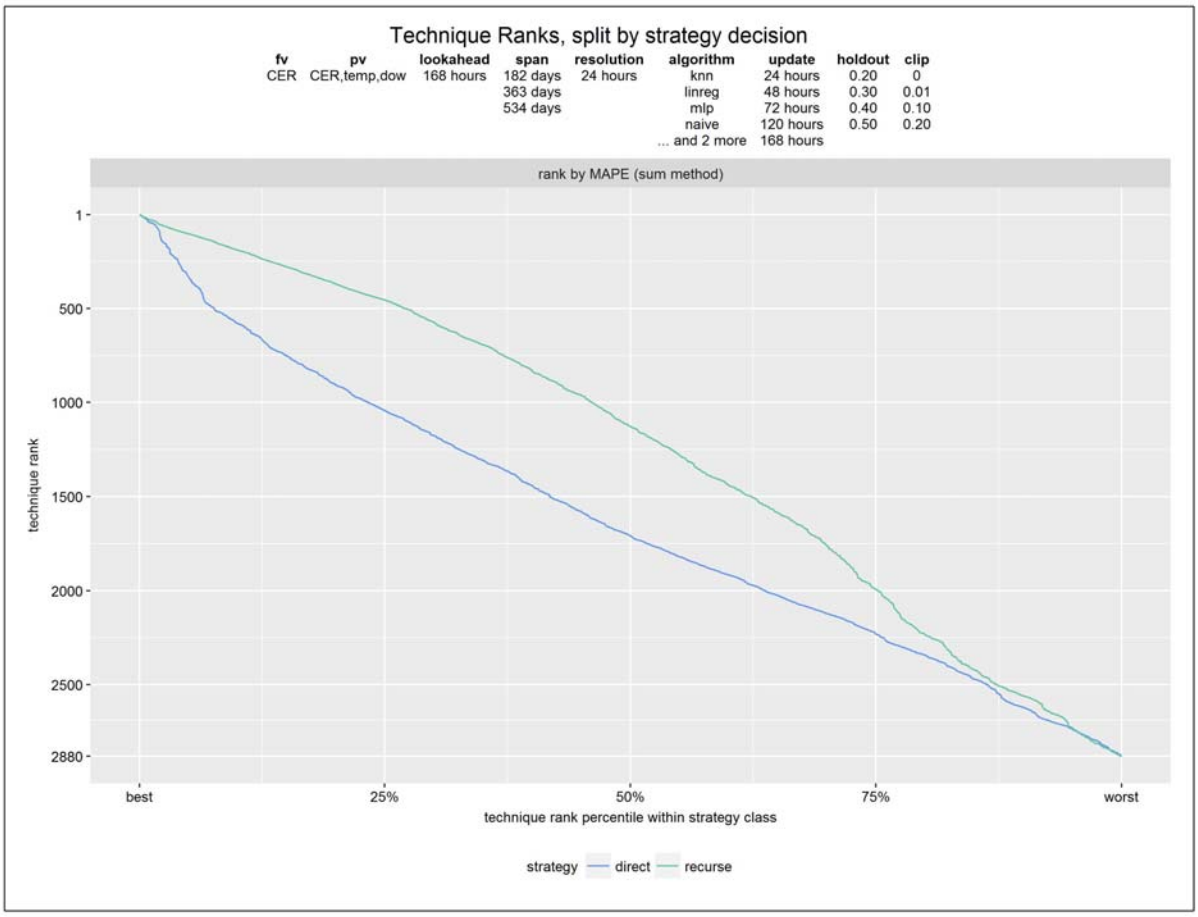


Figure 4-28: Technique rank trend, split by extension rule, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the extension rule decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the extension rule decision option.

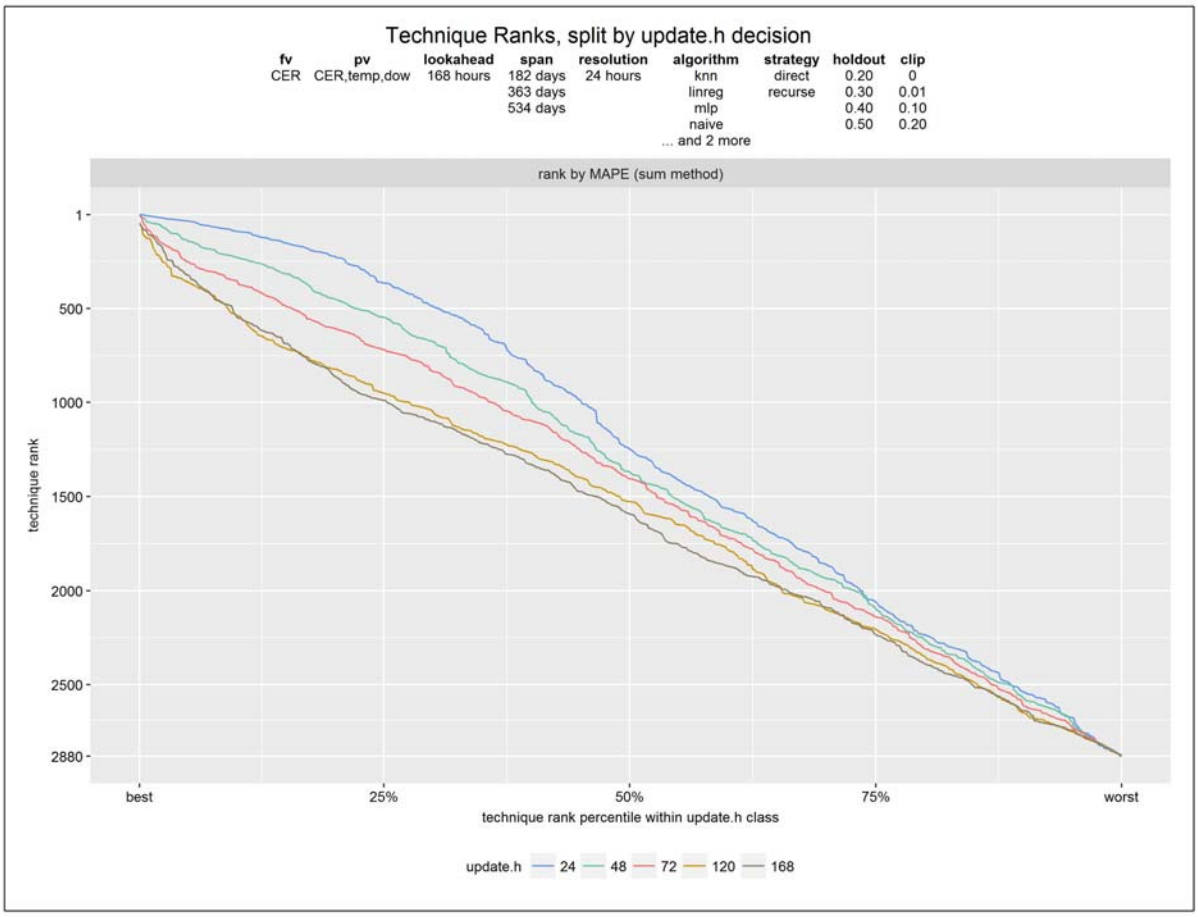


Figure 4-29: Technique rank trend, split by update cycle, Ireland, week-ahead forecast. 2,880 techniques and forecasts. Each curve represents a family of techniques with the update cycle decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the update cycle decision option.

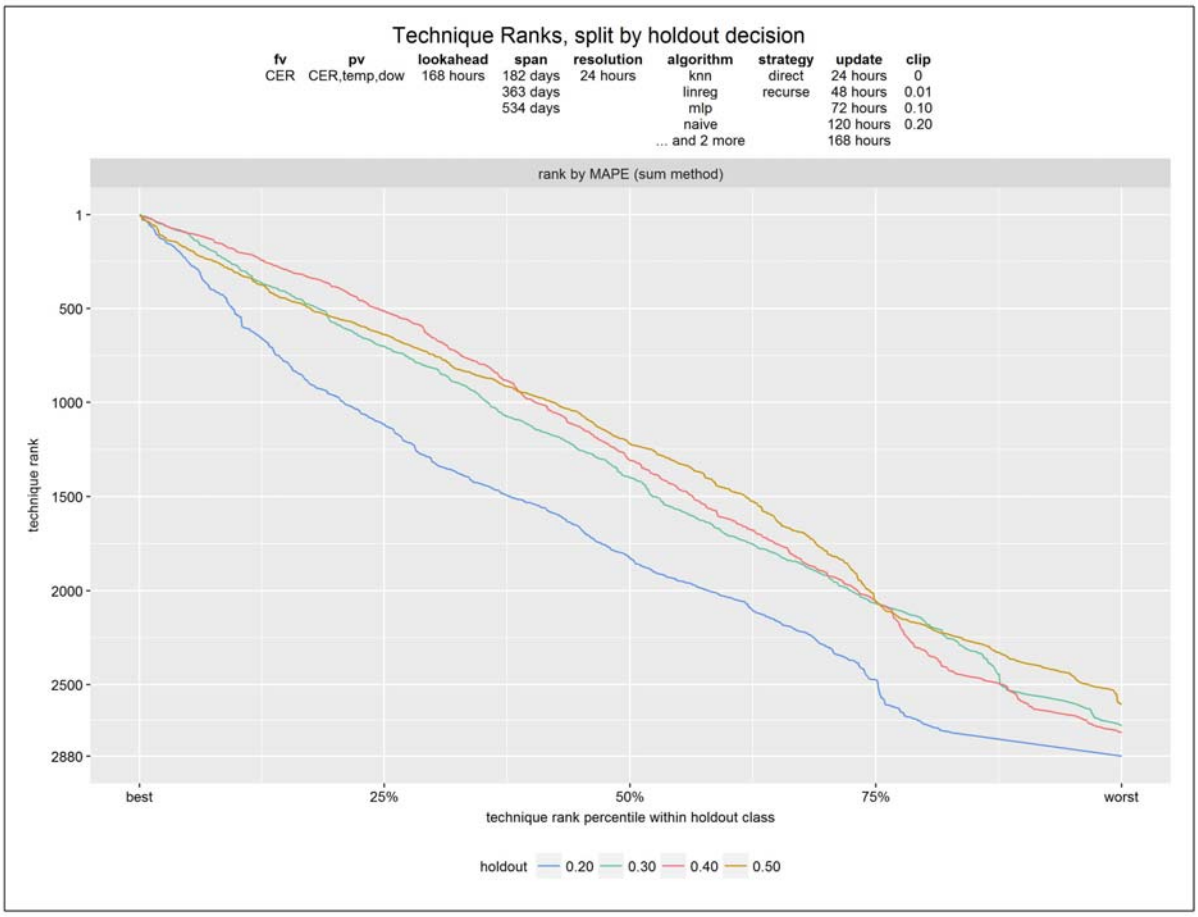


Figure 4-30: Technique rank trend, split by holdout, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the holdout decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the holdout decision option.

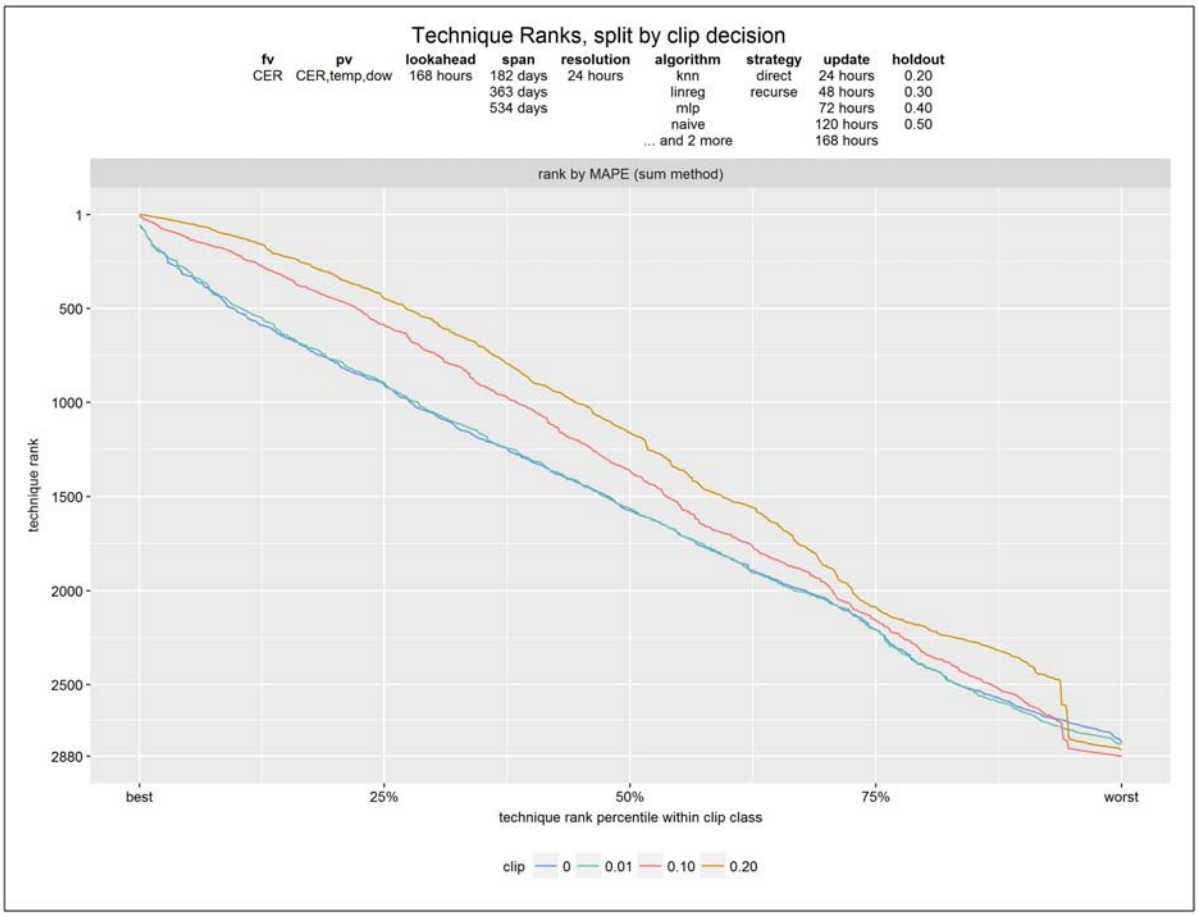


Figure 4-31: Technique rank trend, split by clip, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the clip decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the clip decision option.

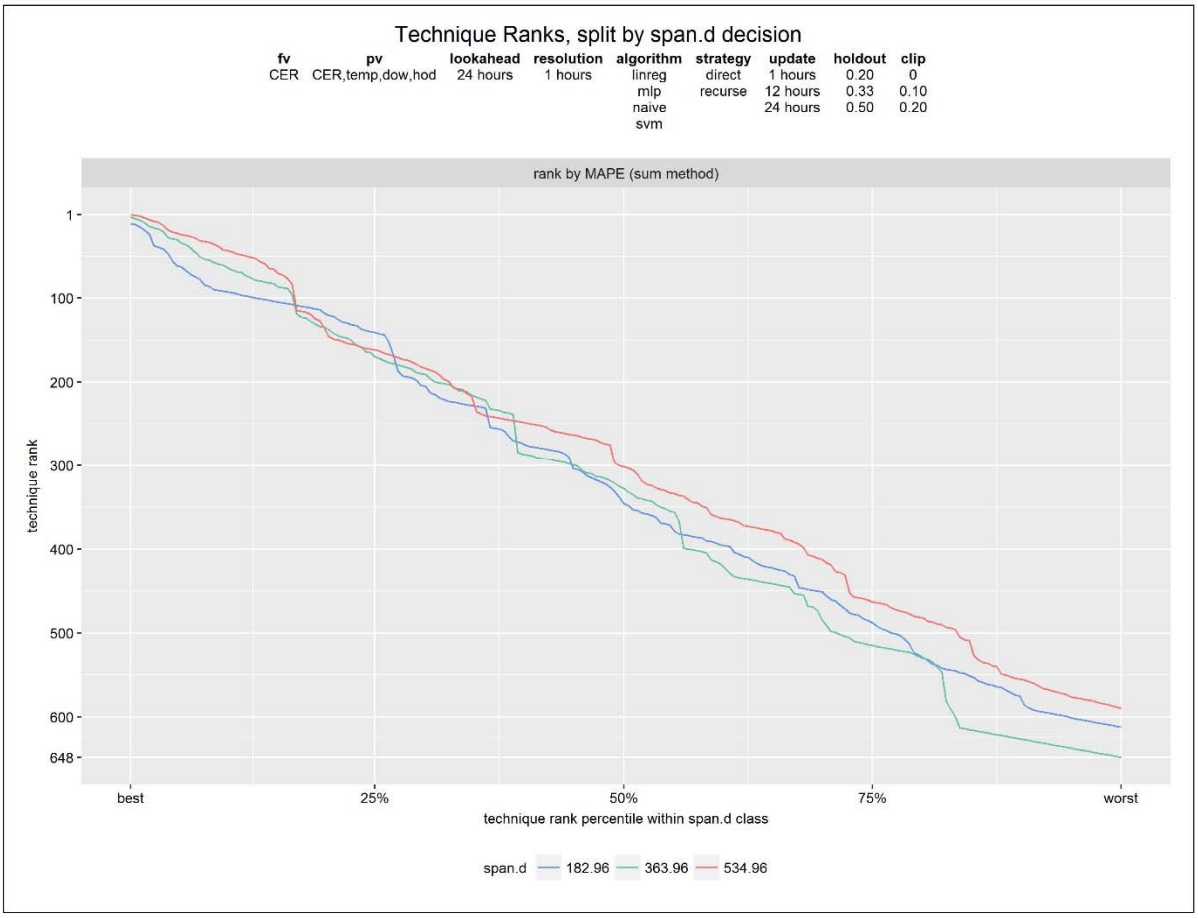


Figure 4-32: Technique rank trend, split by span, Ireland, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the span decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the span decision option.

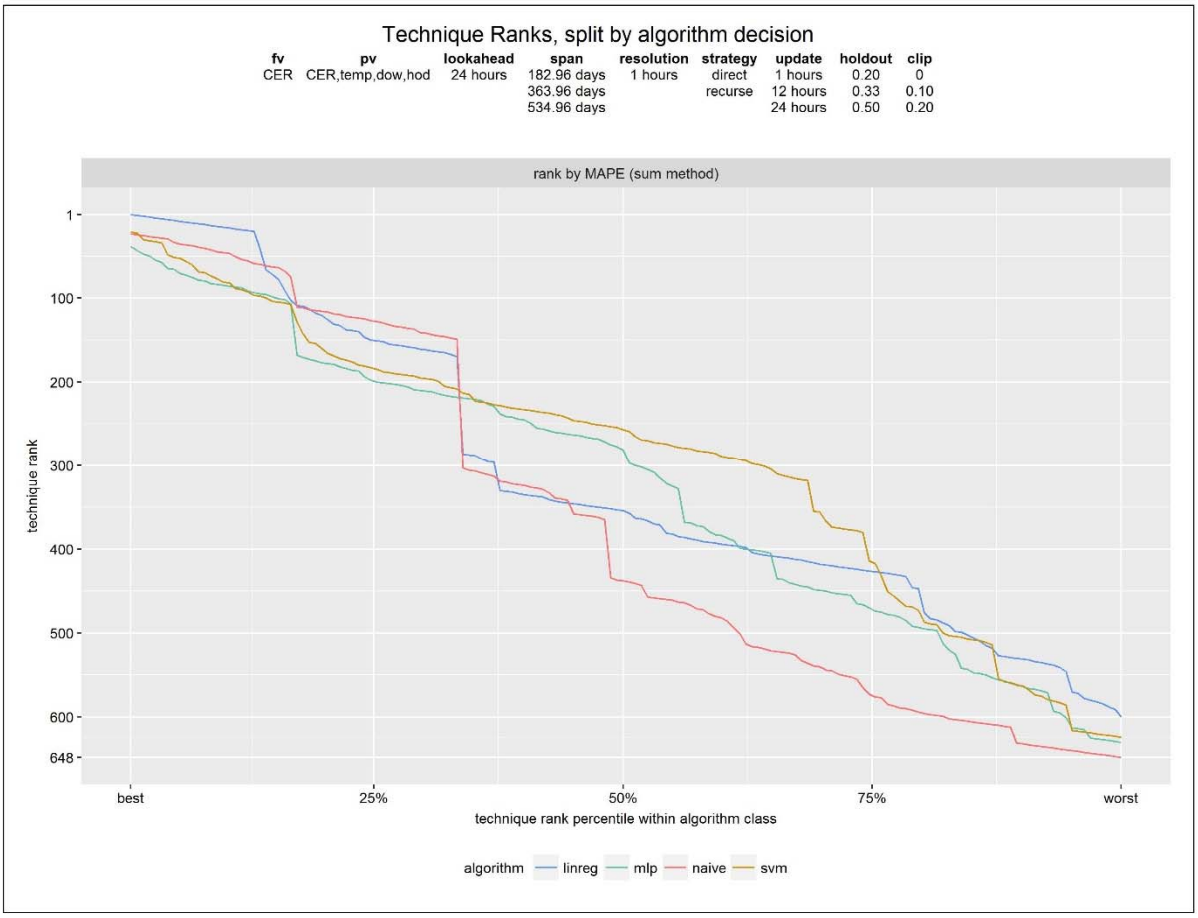


Figure 4-33: Technique rank trend, split by algorithm class, Ireland, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

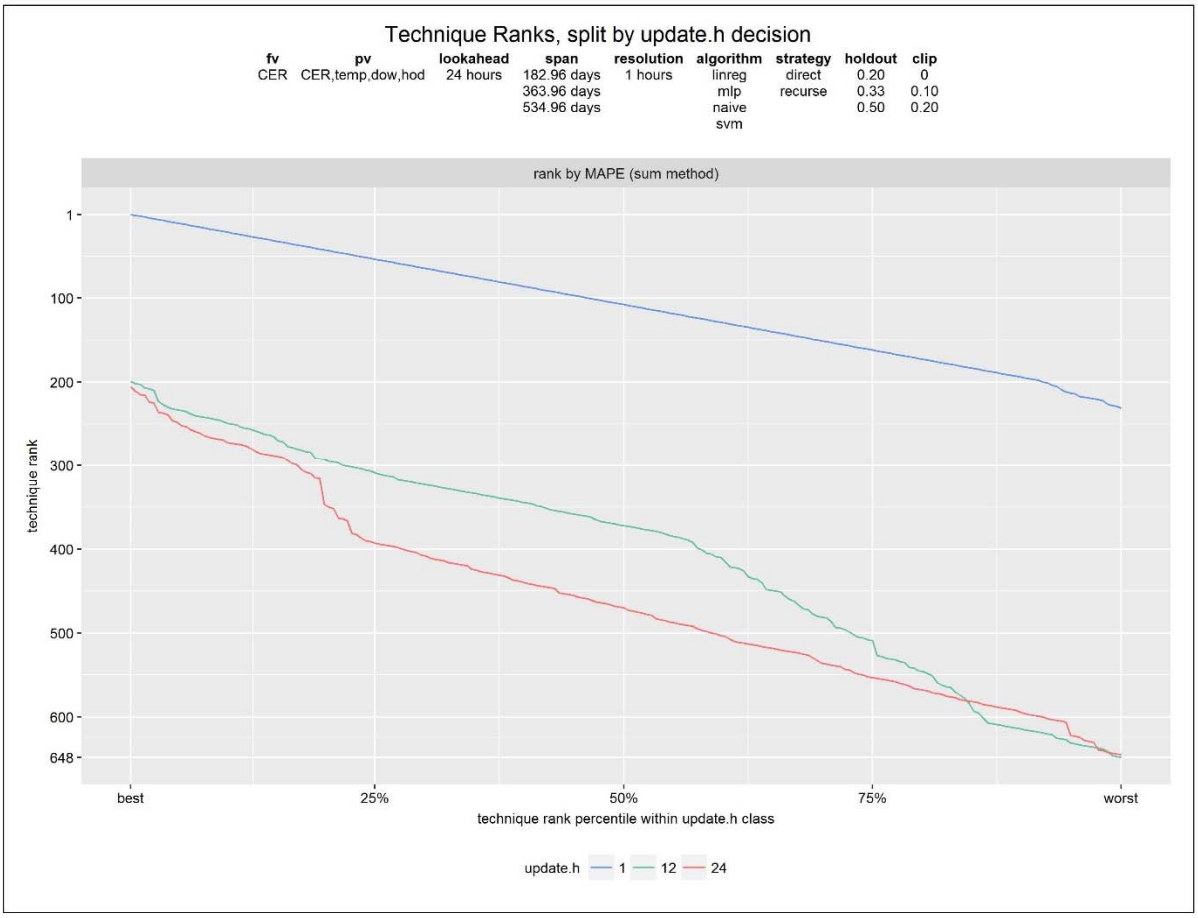


Figure 4-34: Technique rank trend, split by update cycle, Ireland, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the update cycle decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the update cycle decision option.

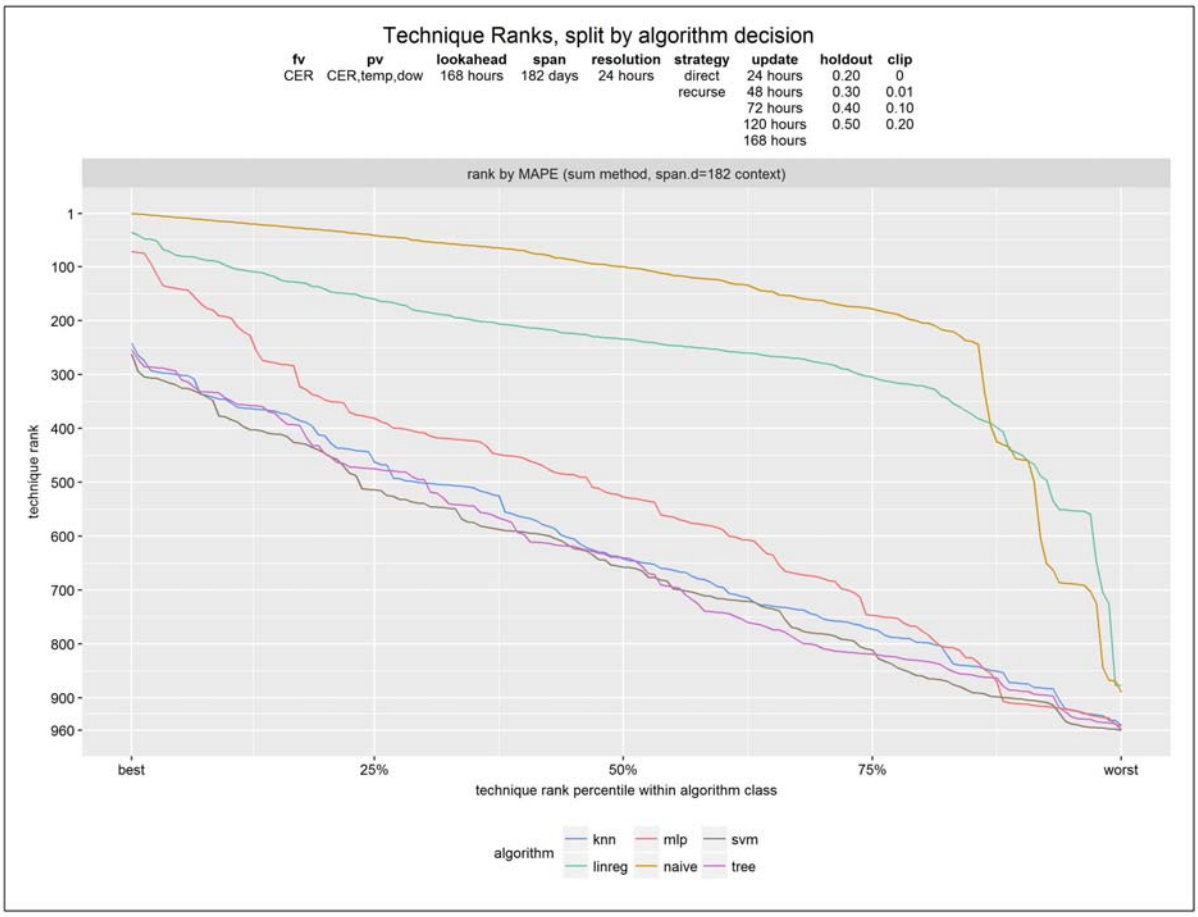


Figure 4-35: Technique rank trend at short span, split by algorithm class, Ireland, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 182-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

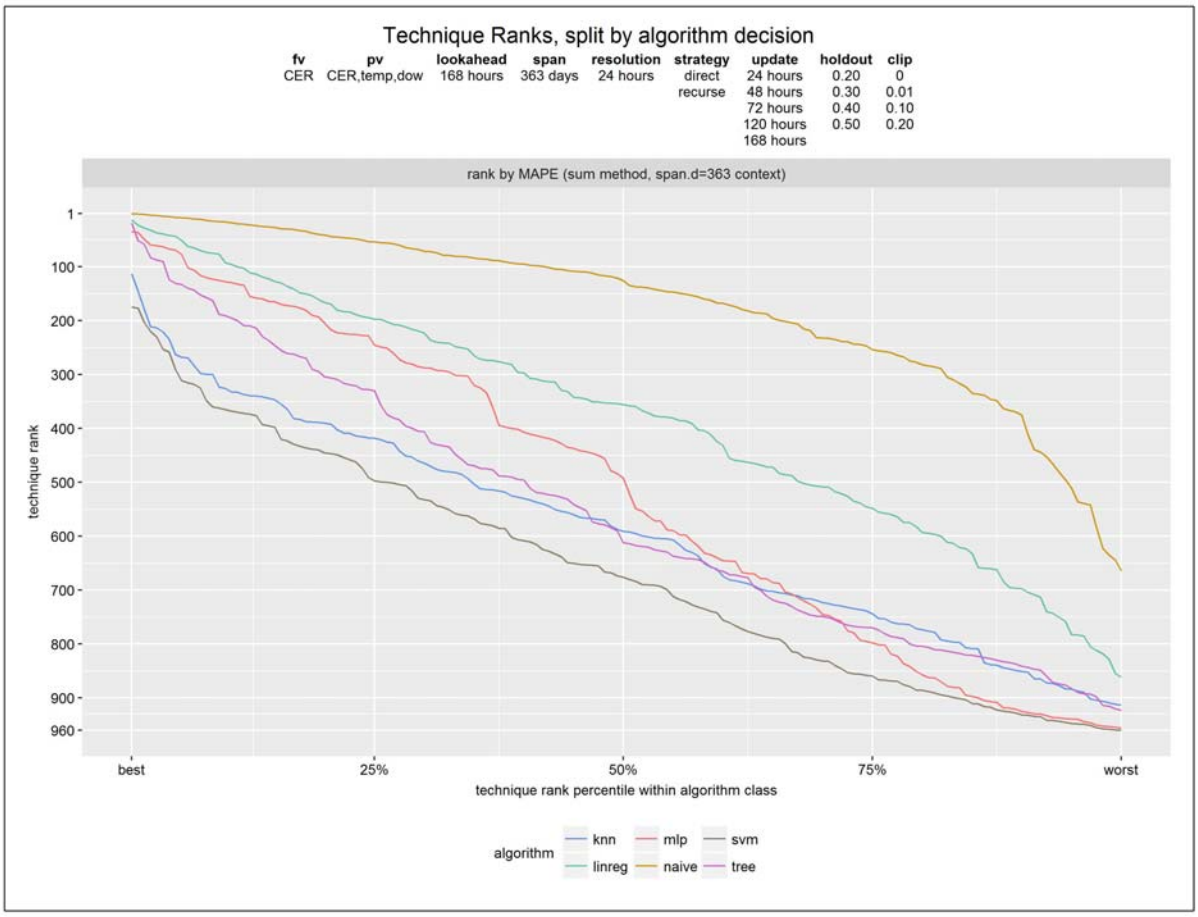


Figure 4-36: Technique rank trend at medium span, split by algorithm class, Ireland, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 363-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

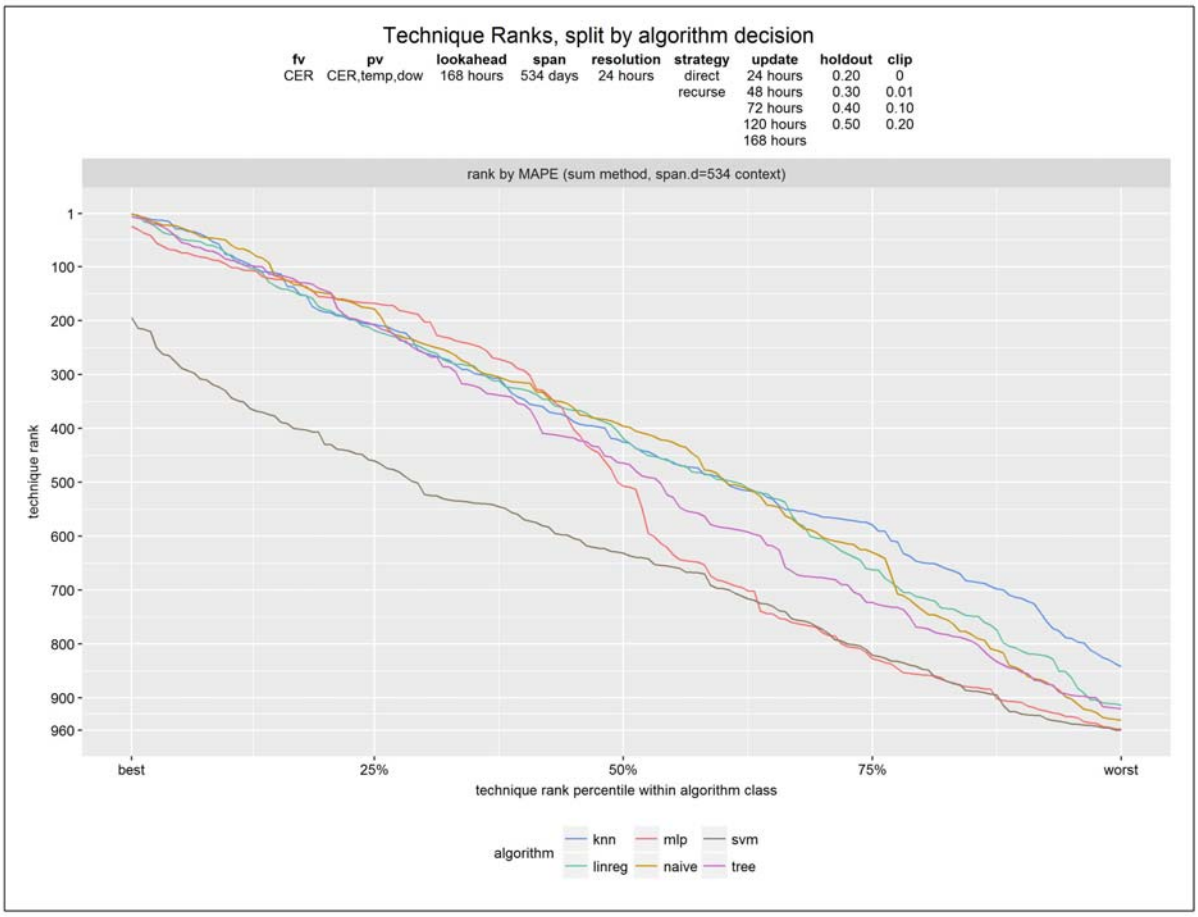


Figure 4-37: Technique rank trend at long span, split by algorithm class, Ireland, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 534-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

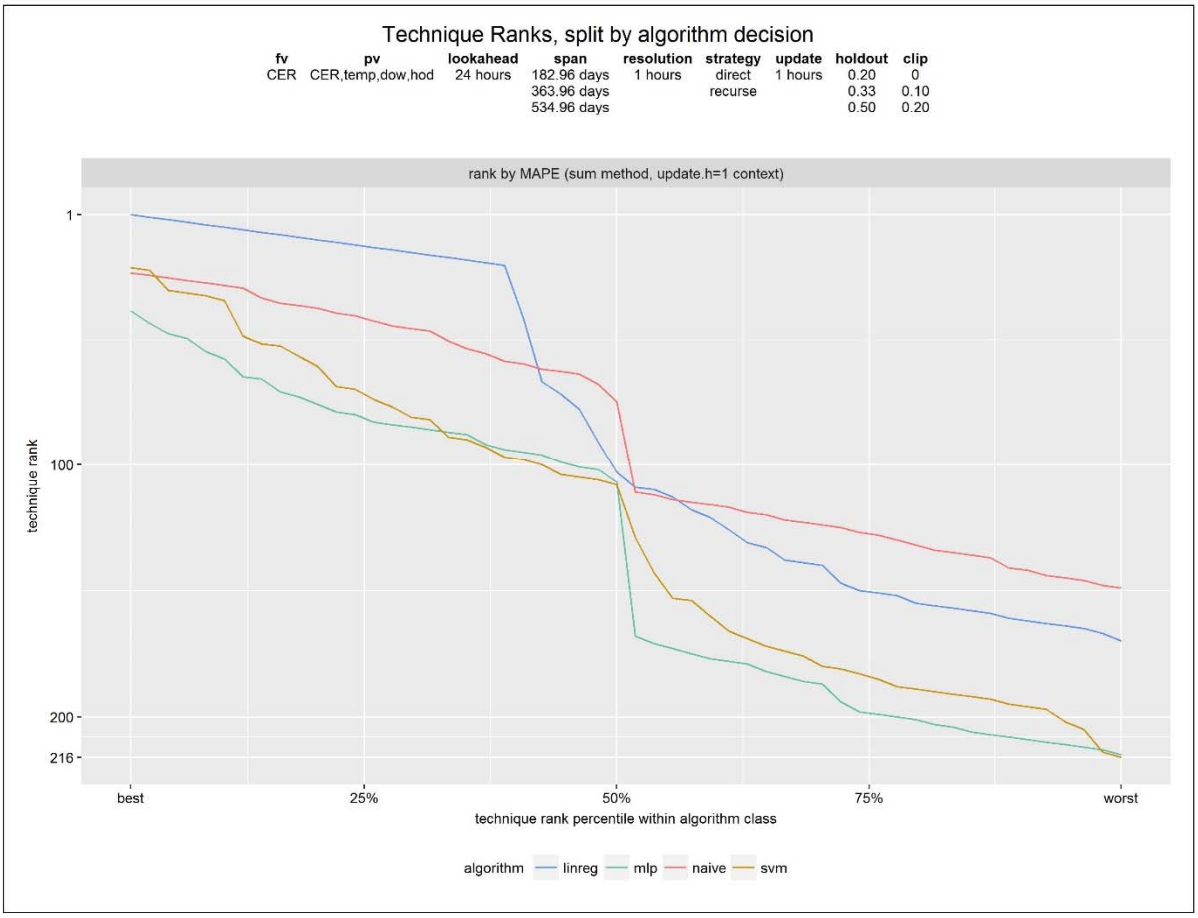


Figure 4-38: Technique rank trend at short update cycle, split by algorithm class, Ireland, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 1-hour update cycle. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

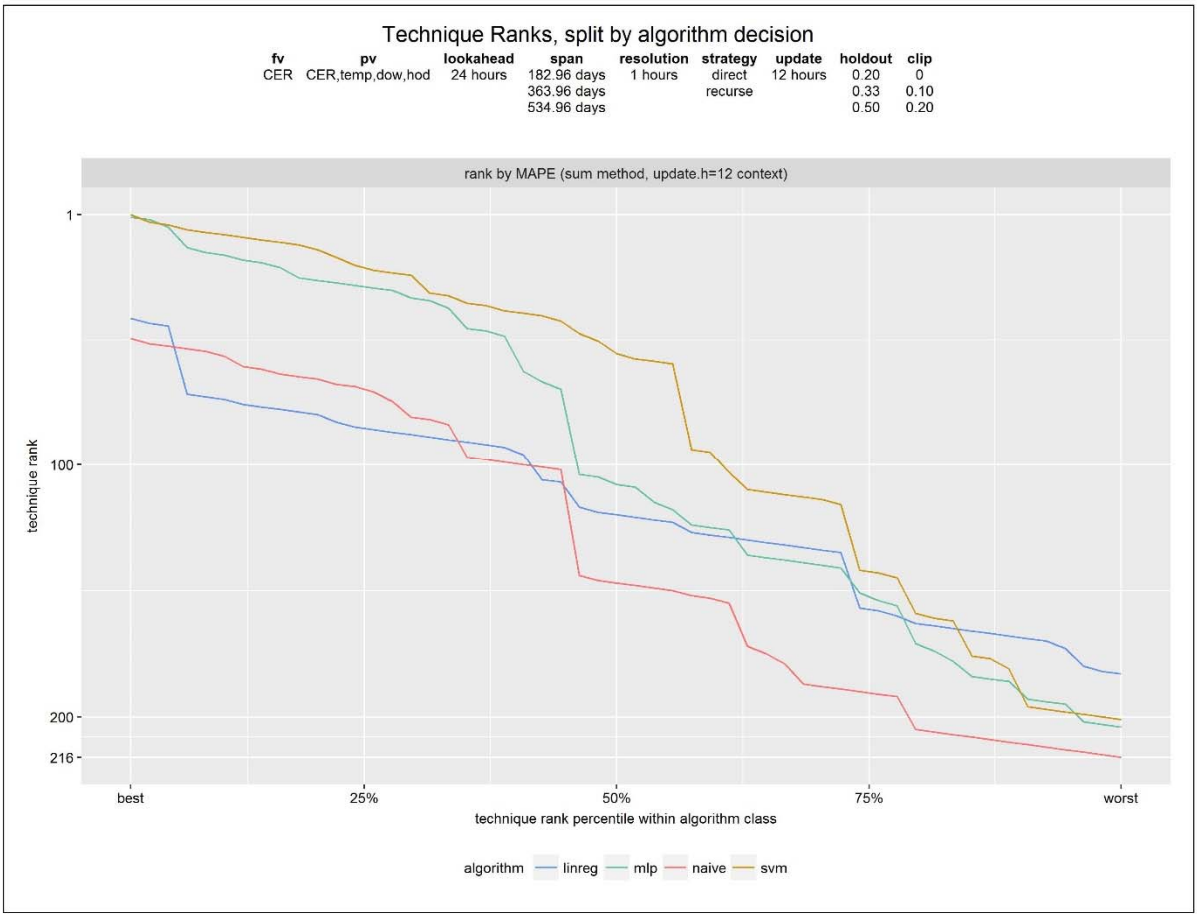


Figure 4-39: Technique rank trend at medium update cycle, split by algorithm class, Ireland, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 12-hour update cycle. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

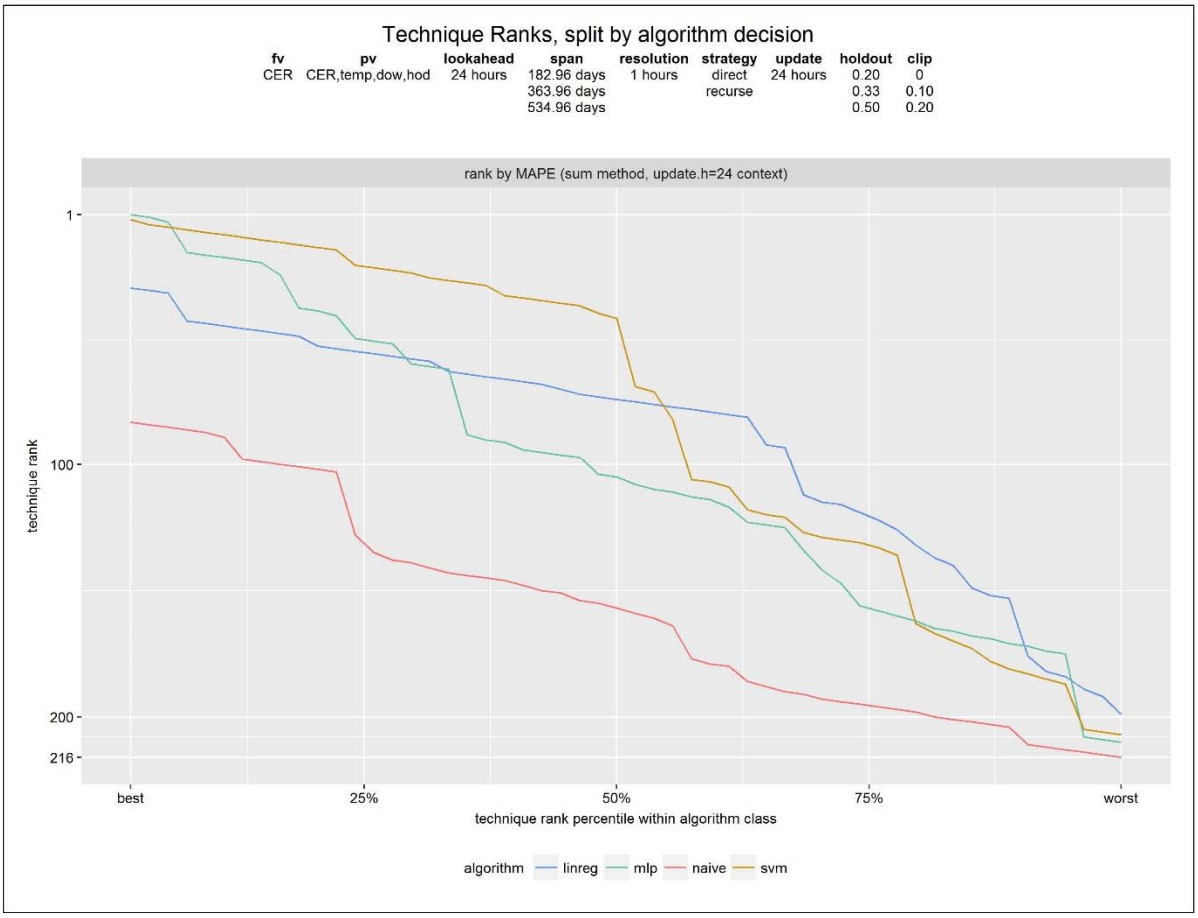


Figure 4-40: Technique rank trend at medium update cycle, split by algorithm class, Ireland, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 24-hour update cycle. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

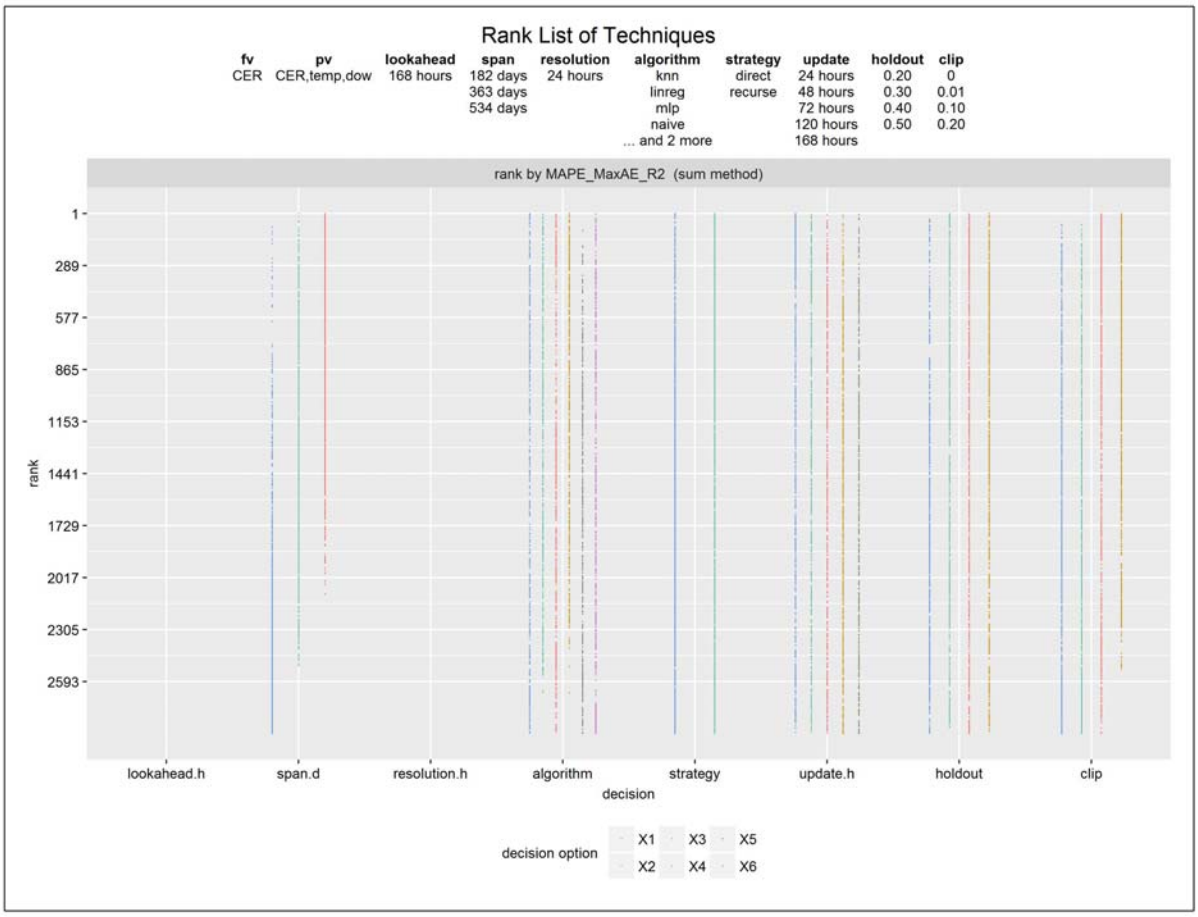


Figure 4-41: List of techniques, rank ordered by performance per MAPE, MaxAE, and R2 scores, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

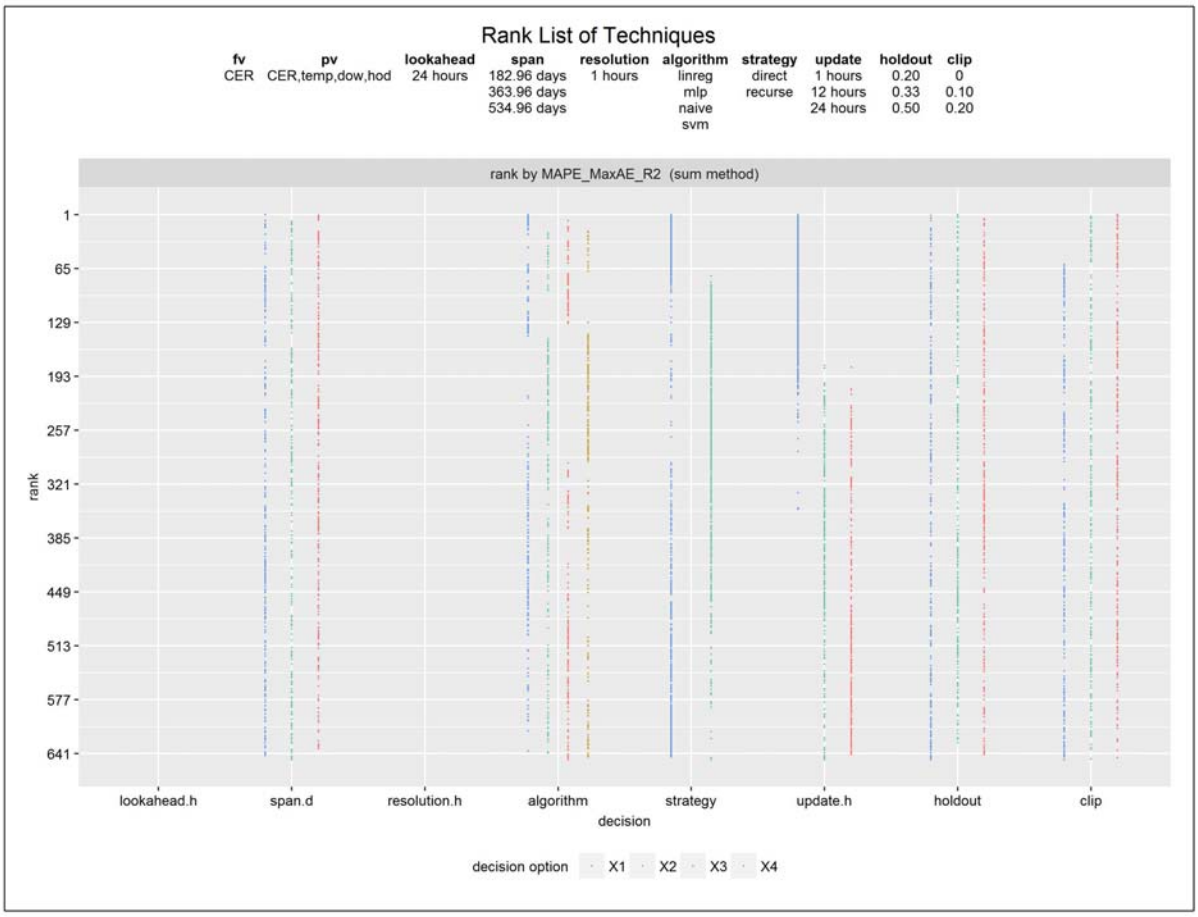


Figure 4-42: List of techniques, rank ordered by performance per MAPE, MaxAE, and R2 scores, Ireland, day-ahead forecasts. 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

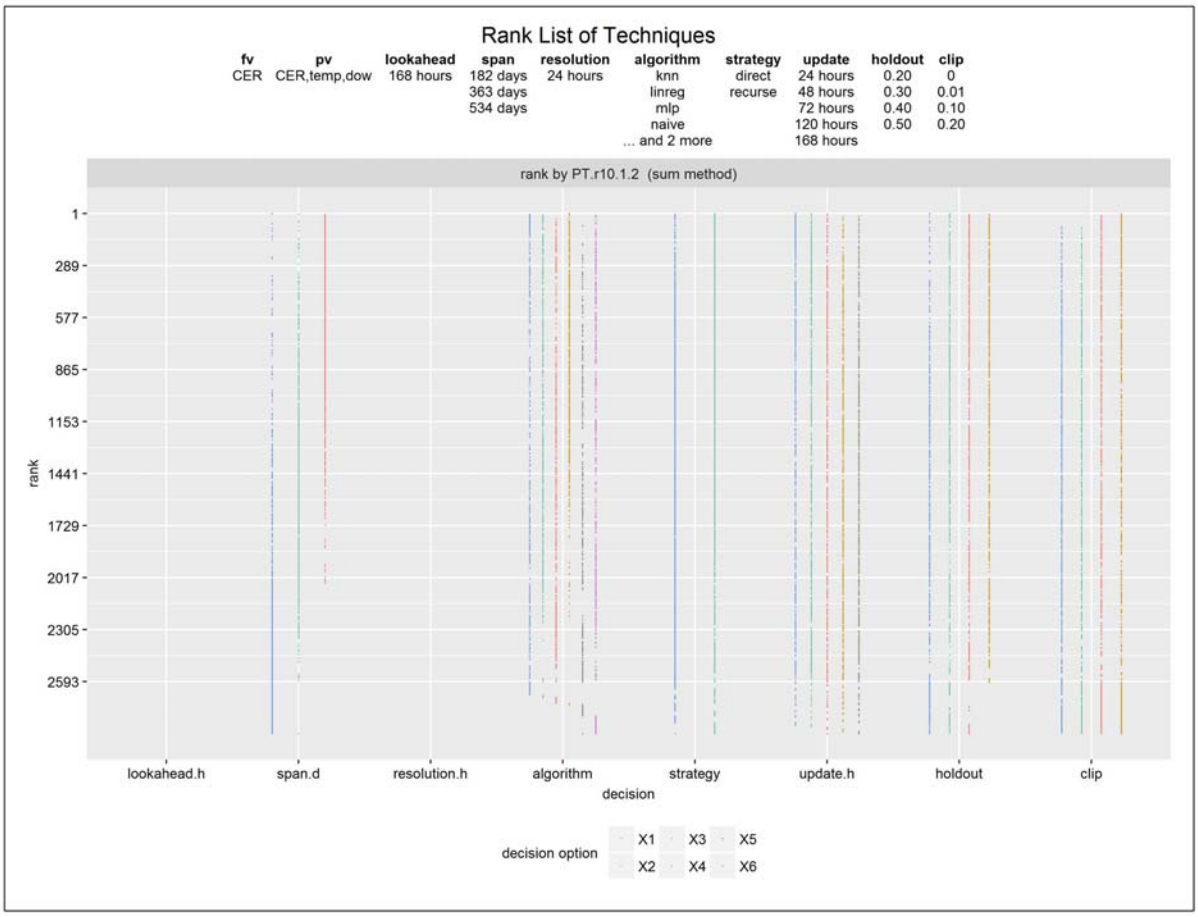


Figure 4-43: List of techniques, rank ordered by performance per PT.r10.1.2 penalty function score, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

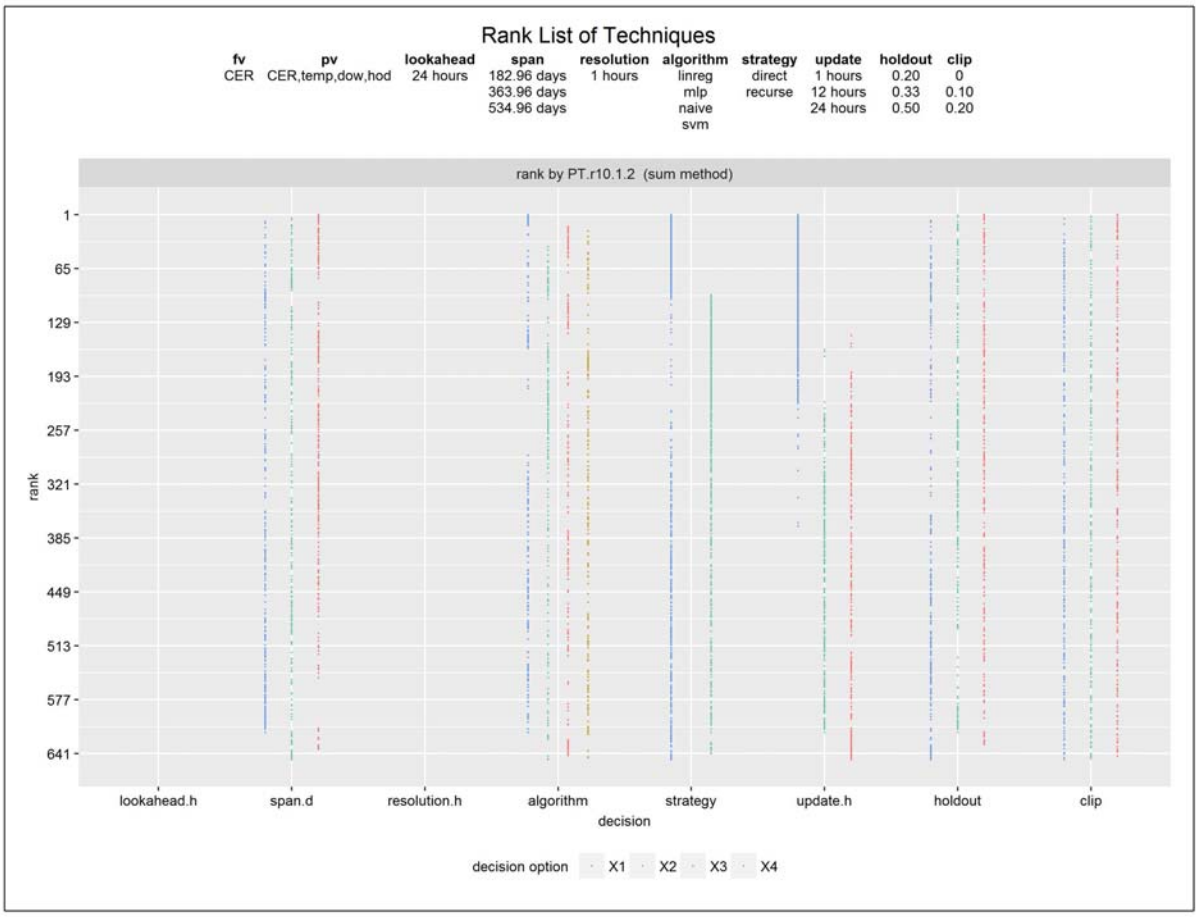


Figure 4-44: List of techniques, rank ordered by performance per PT.r10.1.2 penalty function score, Ireland, day-ahead forecasts. 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

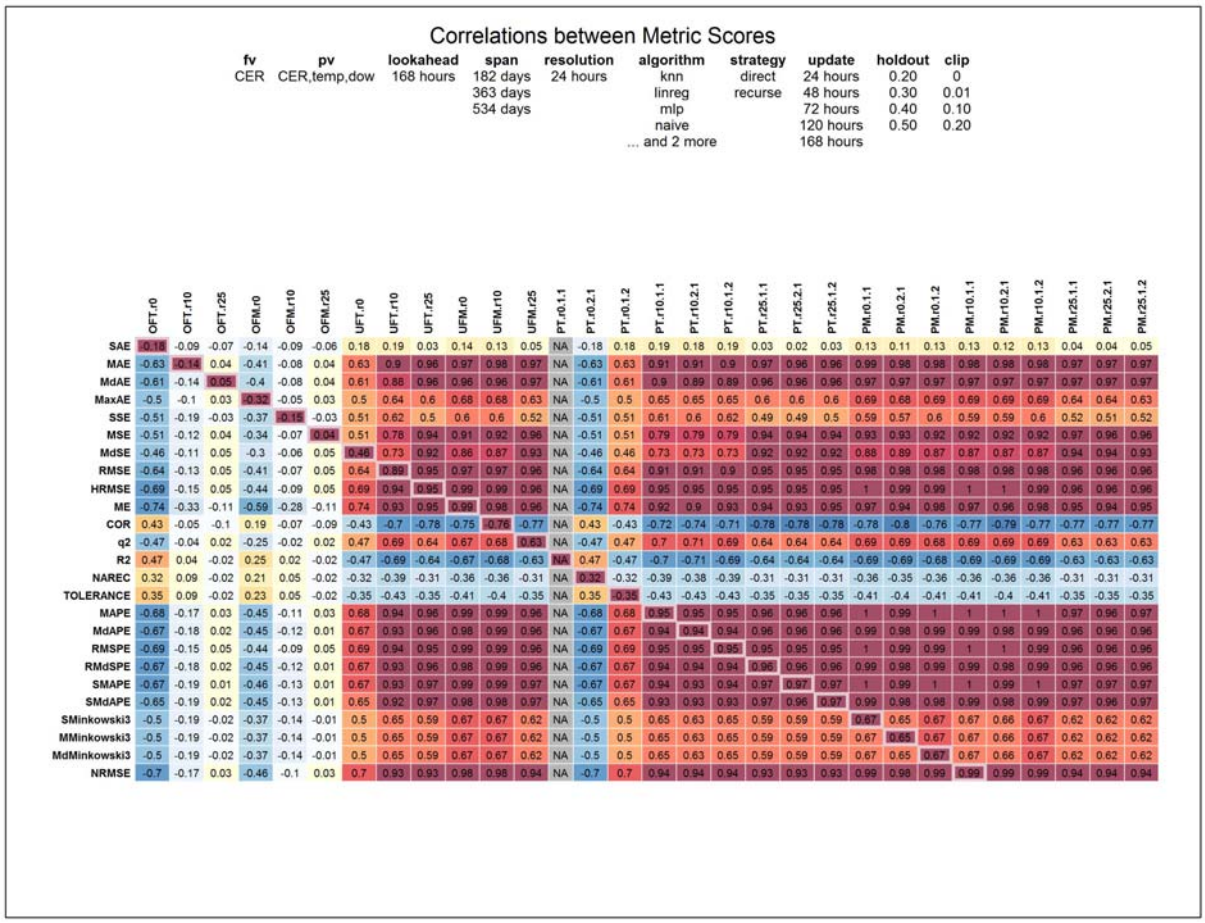


Figure 4-45: Correlations of scores per one standard metric-to-scores per one penalty metric, across metric pairs, Ireland, week-ahead forecasts. 2,880 techniques and forecasts. For each cell, all techniques are scored per two metrics, and the sequence of scores per the first metric are correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation.

5 DATA STRATEGY DECISIONS AND RESIDENTIAL ELECTRICITY DEMAND ESTIMATION

"It is a capital mistake to theorize before one has data."

– Arthur Conan Doyle

5.1 Research Questions

In this chapter, we address the following research questions:

- How robust are residential electricity demand forecasting techniques to data strategies?
- How can smart electric grid data be manipulated to improve residential electricity demand forecasting performance?

5.2 Research Approach

Decisions about data strategy are of special interest to forecasting practitioners due to their potential economic implications. By data strategy, we mean how data is manipulated as it is gathered, corresponding to the discretion afforded in a smart electric grid design over the form in which data could be provided. Manipulation can take the form of transforming data to (1) a sample of the data, (2) a group of clusters, where the clusters are an exhaustive and mutually exclusive partitioning of the data, and each cluster is internally relatively homogeneous according to some similarity criterion, or (3) a temporal magnification of the data, where the time step becomes finer or coarser. A data strategy can include one, two, or all three of these manipulations.

Given a baseline for forecasting performance, we apply our analysis methods and computation platform to study 5,704 week-ahead and day-ahead electricity demand forecasting techniques

scored by 60 metrics, utilizing Ireland CER as before, but focusing on decisions about data strategy.

Sample size	Adjusts for the fraction of households used to represent the whole population. Forecasts are constructed for the sample, then re-scaled for evaluation against the actual whole population.
Cluster count	Adjusts for the number of clusters of households. Forecasts are constructed for each cluster separately, then aggregated for evaluation against the actual whole population.
Time step size	Adjusts for the time step size of raw data collected – daily, hourly, half-hourly, etc. Forecasts are constructed using data at a pre-forecast time step size, then coarsened or refined for evaluation against the whole population at the actually measured time step size.

5.3 Scope of Analysis

We scope our analysis to the effects of 4 decisions and a collection of electricity-related datasets. We choose these particular decisions because they are represented in several other studies that consider their effects in isolation, whereas we are interested in their combined effect. We intentionally fix the other decisions whose effects were already explored in our earlier analysis, namely update cycle, span, holdout, and clip.

We are interested in both week-ahead and day-ahead forecasting settings, so we instantiate two versions of the model – one for week-ahead and one for day-ahead – and conduct the analysis in two parts accordingly. Each part is further organized into three sub-analyses to explore the effects of sampling, clustering, and temporal magnification separately.

5.3.1 Model Instantiations for Week-Ahead Forecasting

For the week-ahead part of our analysis, we lock-in a one-day time step size.

Objective Decisions	lock-in	Reference data source	Ireland CER smart-grid
	lock-in	Reference look-ahead	1 week
	lock-in	Update cycle	1 day
	lock-in	Time step size	1 day
Data Strategy Decisions	lock-in	Span	18 months
Algorithm Decisions	6 options	Algorithm class	k-nearest neighbor linear regression multilayer perceptron naïve support vector regression decision tree
	lock-in	Predictor data sources	Ireland CER smart-grid + WU temperature + day of week
	lock-in	Predictor look-backs	(0, 1 week, 2 weeks, 3 weeks) + (0) + (0)
	2 options	Extension rule	direct recurse
Training and Testing Decisions	lock-in	Holdout %	30%
	lock-in	Clip %	0
Metric and Penalty Decisions	60 options	Metric function or penalty function	any of 30 metrics functions or 30 penalty functions

Sampling

Our model instantiation for sampling analysis specifies 732 techniques, scored by 60 metrics, for a total of 43,920 vectors = 61 sample options x 6 algorithm class options x 2 extension rule options x 60 metric options.

Data Strategy Decisions	6 options	Sample size	any of 20 1-household samples any of 10 39-household samples any of 10 78-household samples any of 10 195-household samples any of 10 586-household samples full population

Clustering

Our model instantiation for clustering analysis specifies 60 techniques, scored by 60 metrics, for a total of 3,600 vectors = 5 cluster grouping options x 6 algorithm class options x 2 extension rule options x 60 metric options – or 28,800 component vectors = 2+4+8+25+1 cluster options x 6 algorithm class options x 2 extension options x 60 metric options.

Data Strategy Decisions	5 options	Cluster count	partitioned into 2 clusters partitioned into 4 clusters partitioned into 8 clusters partitioned into 25 clusters not clustered

Temporal Magnification

Our model instantiation for temporal magnification analysis specifies 36 techniques, scored by 60 metrics, for a total of 2,160 vectors = 3 time step size options x 6 algorithm class options x 2 extension rule options x 60 metric options.

Data Strategy Decisions	3 options	Time step size	6 hour 12 hour 1 day

5.3.2 Model Instantiations for Day-Ahead Forecasting

For the week-ahead part of our analysis, we lock-in a one-day time step size.

Objective Decisions	lock-in	Reference data source	Ireland CER smart-grid
	lock-in	Reference look-ahead	1 day
	lock-in	Update cycle	1 hour
	lock-in	Time step size	1 hour
Data Strategy Decisions	lock-in	Span	18 months
Algorithm Decisions	4 options	Algorithm class	linear regression multilayer perceptron naïve support vector regression
	lock-in	Predictor data sources	Ireland CER smart-grid + WU temperature + day of week
	lock-in	Predictor look-backs	(0, 1 day, 2 days, 3 days, 4, days, 5 days, 6 days) + (0) + (0) + (0)
	2 options	Extension rule	direct recurse
Training and Testing Decisions	lock-in	Holdout %	30%
	lock-in	Clip %	0
Metric and Penalty Decisions	60 metrics	Metric function or penalty function	any of 30 metrics functions or 30 penalty functions

Sampling

Our model instantiation for sampling analysis specifies 488 techniques, scored by 60 metrics, for a total of 29,280 vectors = 61 sample options x 4 algorithm class options x 2 extension rule options x 60 metric options.

Data Strategy Decisions	6 options	Sample size	any of 20 1-household samples any of 10 39-household samples any of 10 78-household samples any of 10 195-household samples any of 10 586-household samples full population

Clustering

Our model instantiation for clustering analysis specifies 40 techniques, scored by 60 metrics, for a total of 2,400 vectors = 5 cluster grouping options x 4 algorithm class options x 2 extension rule options x 60 metric options – or 19,200 component vectors = 2+4+8+25+1 cluster options x 4 algorithm class options x 2 extension options x 60 metric options.

Data Strategy Decisions	5 options	Cluster count	partitioned into 2 clusters partitioned into 4 clusters partitioned into 8 clusters partitioned into 25 clusters not partitioned

Temporal Magnification

Our model instantiation for temporal magnification analysis specifies 24 techniques, scored by 60 metrics, for a total of 1,440 vectors = 3 time step size options x 4 algorithm class options x 2 extension rule options x 60 metric options.

Data Strategy Decisions	3 options	Time step size	30 min 1 hour 2 hours

5.3.3 Data Sources

We specify reference series and predictor series as renderings of data from the following sources, as before.

Reference series	Electricity usage. From Ireland Commission for Energy Regulation Smart Metering Project (CER). 30-minute intervals across 782 control households in Dublin, Ireland in 2009-10.
Predictor series	Electricity usage. From Ireland Commission for Energy Regulation Smart Metering Project (CER). Measured at 30-minute intervals across 782 households in Dublin, Ireland in 2009-10. *Same data source as for reference series.* Temperature. From IBM Weather Underground. Measured at 1-hour intervals at Dublin, Ireland in 2009-10. Day of week. Generated by analysis platform. Hour of day (for day-ahead only). Generated by analysis platform.

5.4 Results

For expediency and because MAPE is a de facto electricity forecasting standard, as noted earlier, we highlight our results around MAPE-based techniques.

5.4.1 Sampling

Some electricity usage series, aggregated over samples, resemble the series for the population, while others appear very different. Consequently, forecasts based on techniques that use the series, when re-scaled to population size, also vary accordingly (Figure 5-6, Figure 5-7). Some severely under- or over-forecast at the population scale, some do both.

With both week-ahead and day-ahead forecasts, MAPE scores vary widely across techniques applied to single household samples, but as sample size increases to only 5%-10% of population size, they converge quickly to tight distributions approximating the population distribution (Figure 5-8, Figure 5-9).

Technique ranks within a single household sample vary widely across the set of all single household samples – techniques that perform well for one household do not generally perform

very well for other households (Figure 5-10, Figure 5-16). As sample size increases, technique ranks within a sample, as a group, move monotonically to align with technique ranks within the population, though not as quickly as MAPE score moves (Figure 5-11, Figure 5-12, Figure 5-13, Figure 5-14, Figure 5-15, Figure 5-17). Notably, however, individual technique ranks within a sample converge on the population rank, but not monotonically, i.e., for any specific technique, its mean rank within a sample at some size does not necessarily move closer with each increase in size on its way to the mean rank within the population.

With both week-ahead forecasts and day-ahead forecasts, penalty function scores, e.g., PT.r10.1.2 scores, show a relationship to sample size similar to that seen for MAPE scores, though converging to the population distribution more slowly (Figure 5-18, Figure 5-19).

5.4.2 Clustering

With both week-ahead and day-ahead forecasts, the number of clusters used to produce population forecasts has little effect on MAPE score distributions across techniques (Figure 5-20, Figure 5-21). As number of clusters decreases, technique ranks within a cluster grouping, as a group, move monotonically to align with technique ranks within the population, but individual technique ranks do so non-monotonically (Figure 5-22, Figure 5-23, Figure 5-24, Figure 5-25, Figure 5-26).

5.4.3 Temporal Magnification

With week-ahead forecasts, MAPE score distributions across techniques at finer time step sizes – 6-hour and 12-hour as compared to a 1-day baseline – tend to worsen (Figure 5-27).

With day-ahead forecasts, MAPE score distributions across techniques at finer time step sizes – 30-minute and 2-hour as compared to a 1-hour baseline – tend to improve (Figure 5-28).

Technique ranks show no clear relationship to the time step size used to produce population forecasts (Figure 5-29, Figure 5-30, Figure 5-31, Figure 5-32).

5.5 Insights

We glean the following insights from our results, with the requisite caveat that they are based on one specific smart electric grid dataset and a practically scoped set of experiments.

5.5.1 Sampling

> Forecasting performance is not much degraded by sampling.

We see that use of even a small sample size, say anything larger than 5%-10% of population size, does not much degrade forecasting performance.

5.5.2 Clustering

> Forecasting performance is not much improved by clustering.

We see that none of a wide variety of cluster counts does much to improve forecasting performance.

5.5.3 Temporal Magnification

> Forecasting performance is degraded by refining temporal magnification at coarse time step sizes.

> Forecasting performance is improved by refining temporal magnification at fine time step sizes.

With week-ahead forecasts, we see that even a small decrease in time step size degrades forecasting performance, over the range 6 hours to 24 hours. With day-ahead forecasts, we

see that even a small decrease in time step size improves forecasting performance, over the range 30 minutes to 2 hours.

5.6 Implications for Smart Electric Grid Design and Electricity Policy

Data strategies are related to the economic costs of smart electric grid operation and implementation. As a first order approximation, a penalty function can serve as a proxy cost function, assuming forecast asymmetric error is proportional to the combined impact of all cost elements. A better approximation of economic costs requires a cost function defined in terms of all cost elements or their dependencies. With such a cost function, a distribution of economic costs can be calculated across various potential actual demand levels through Monte Carlo simulation or other methods. In turn, an economic cost estimate or optimal economic cost boundary can be calculated.

Formulation of a cost function is out of scope here, as it requires separate research to uncover the relationships between many cost elements not related to forecasting. [132] However, we do suggest the form that the cost function may take, to expose how forecasting process decisions about data strategy can influence economic cost, and how they are therefore important considerations for smart electric grid design and electricity policy (Figure 5-1).

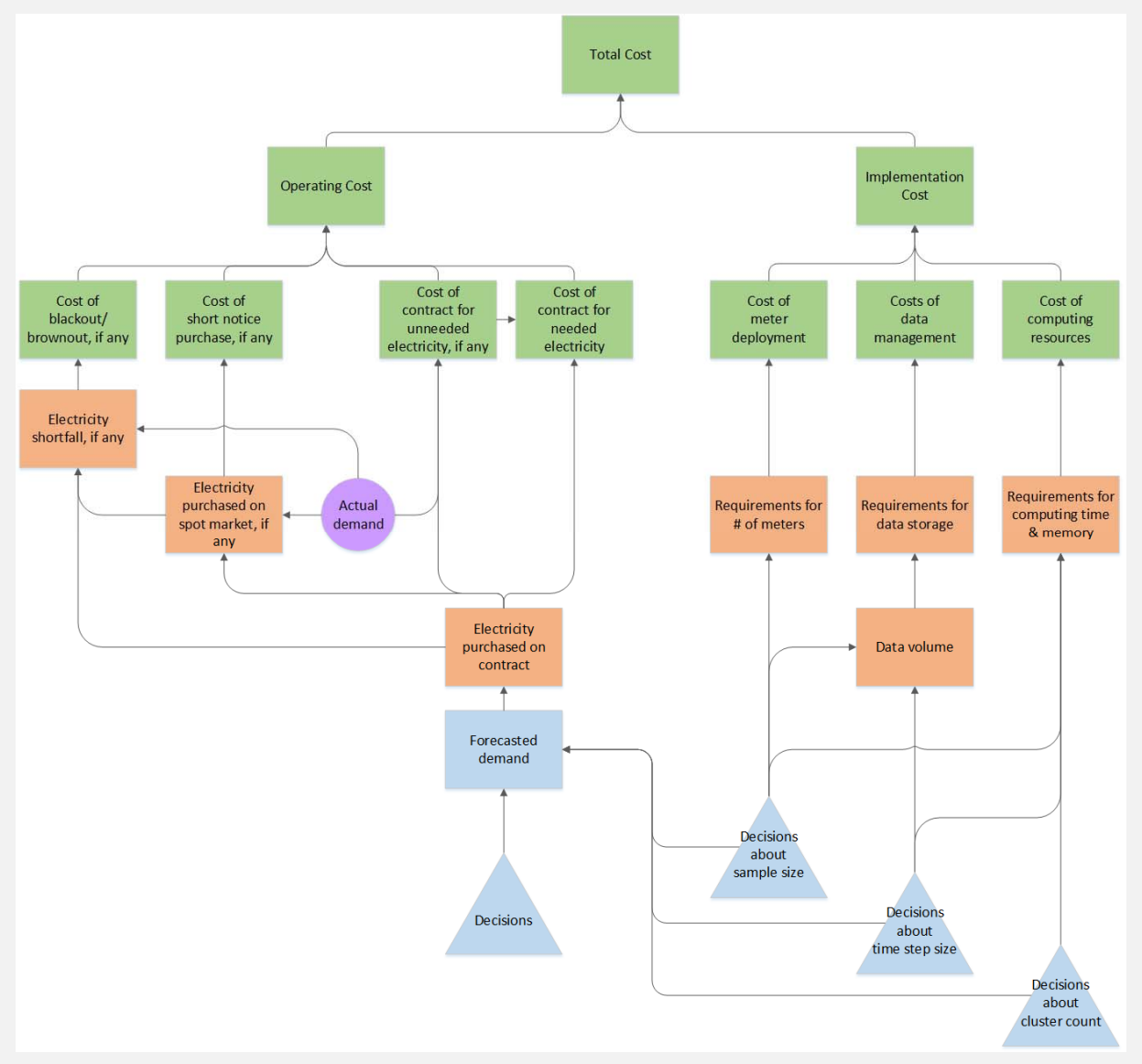
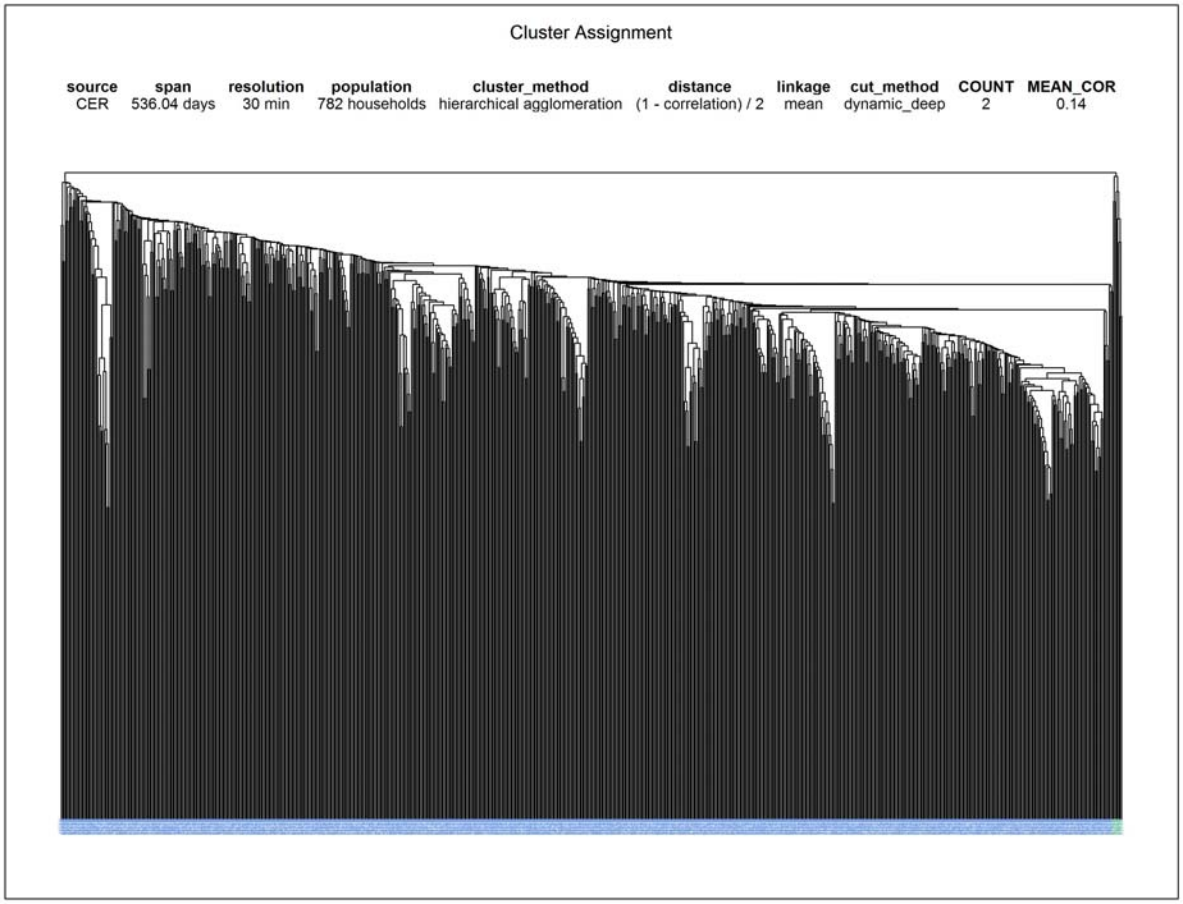


Figure 5-1: Influence diagram of cost function, accounting for decisions about data strategy.

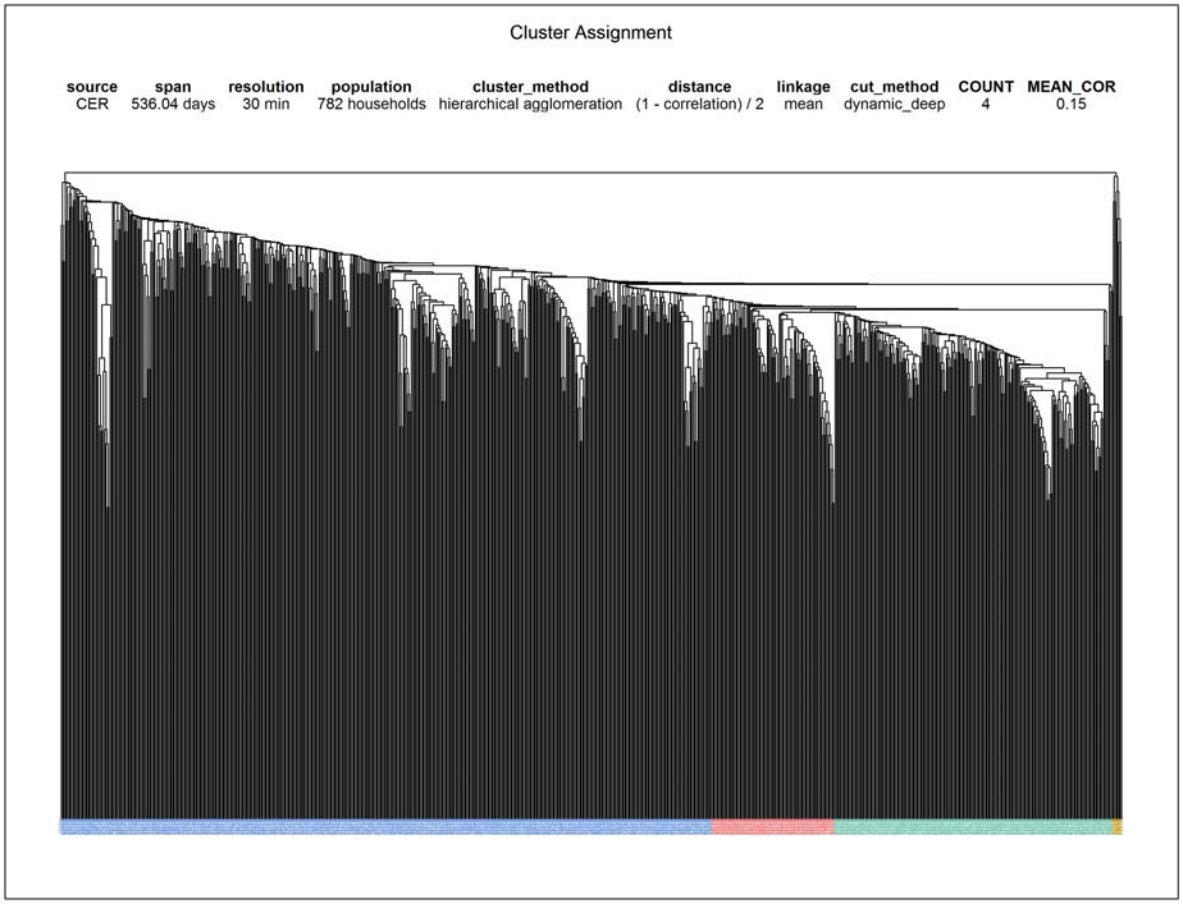
Total cost is a function of …	• Implementation cost • Operating cost
Implementation cost is a function of …	• Cost of meter deployment • Cost of data management • Cost of computing resources
Operating costs is a function (indirectly) of …	• Forecasted demand
Cost of meter deployment is a function of …	• Requirements for # of meters
Cost of data management is a function of …	• Requirements for data storage
Cost of computing resources is a function of …	• Requirements for computing time and memory
Requirements for # of meters is a function of …	• Decision about sample size
Requirements for data storage is a function of …	• Data volume
Requirements for computing time and memory is a function of …	• Decision about sample size • Decision about time step size • Decision about cluster count
Data volume is a function of …	• Decision about sample size • Decision about time step size
Forecasted demand is a function of …	• Decision about sample size • Decision about time step size • Decision about cluster count • Other decisions
Decision about sample size is a decision made by forecasting practitioner	
Decision about time step size is a decision made by forecasting practitioner	
Decision about cluster count is a decision made by forecasting practitioner	

5.7 Tables and Data Visualizations



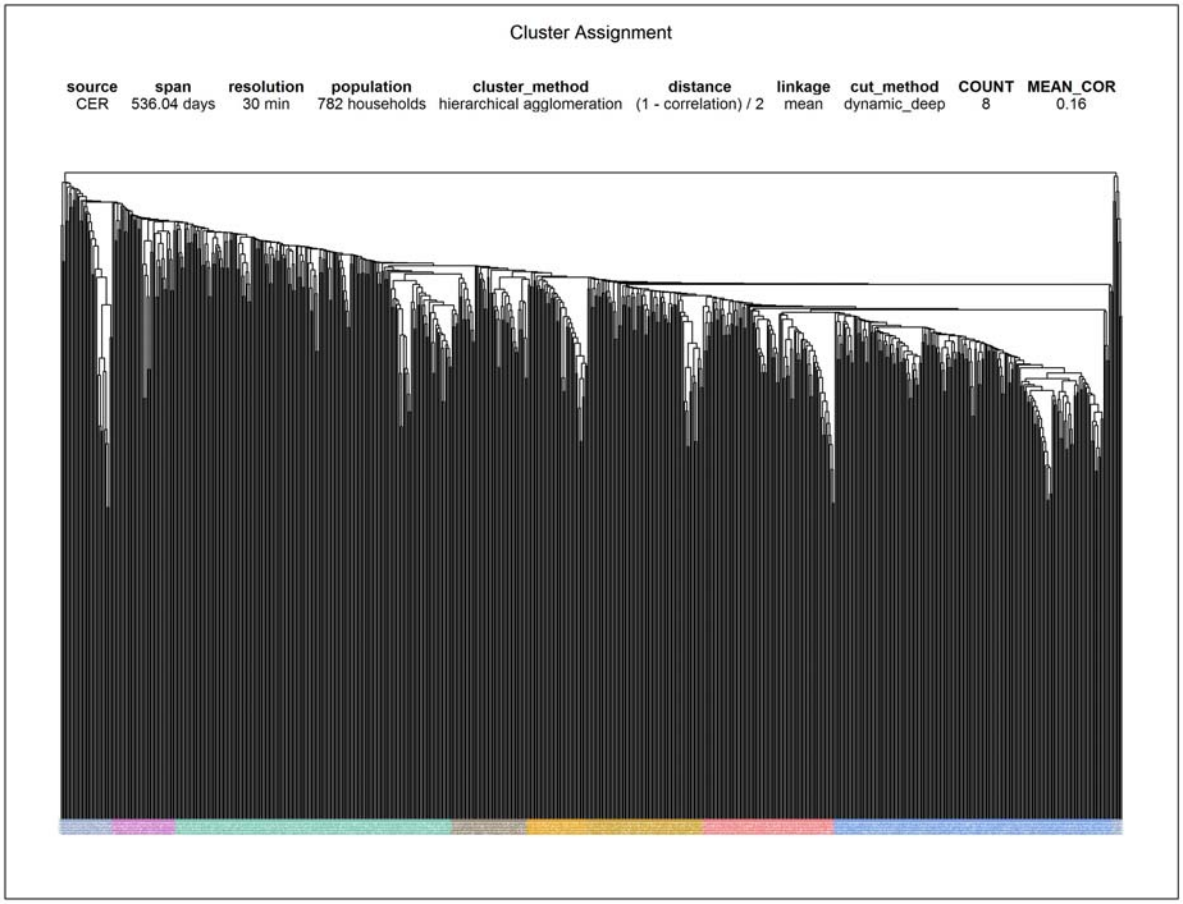

```
2 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  782 (distance mean 0.431, sd 0.038, max 0.672, min 0.239  :::  cor mean 0.138, sd 0.075, max 0.522, min -0.343)
2 clusters of sizes ...
  775 (distance mean 0.430, sd 0.037, max 0.672, min 0.239  :::  cor mean 0.140, sd 0.074, max 0.522, min -0.343)
    7 (distance mean 0.474, sd 0.028, max 0.505, min 0.385  :::  cor mean 0.051, sd 0.055, max 0.230, min -0.009)
                                                            weighted mean 0.140
```

Figure 5-2: Group 782 households as 2 clusters, Ireland. *Colors* indicate to which clusters households are assigned.



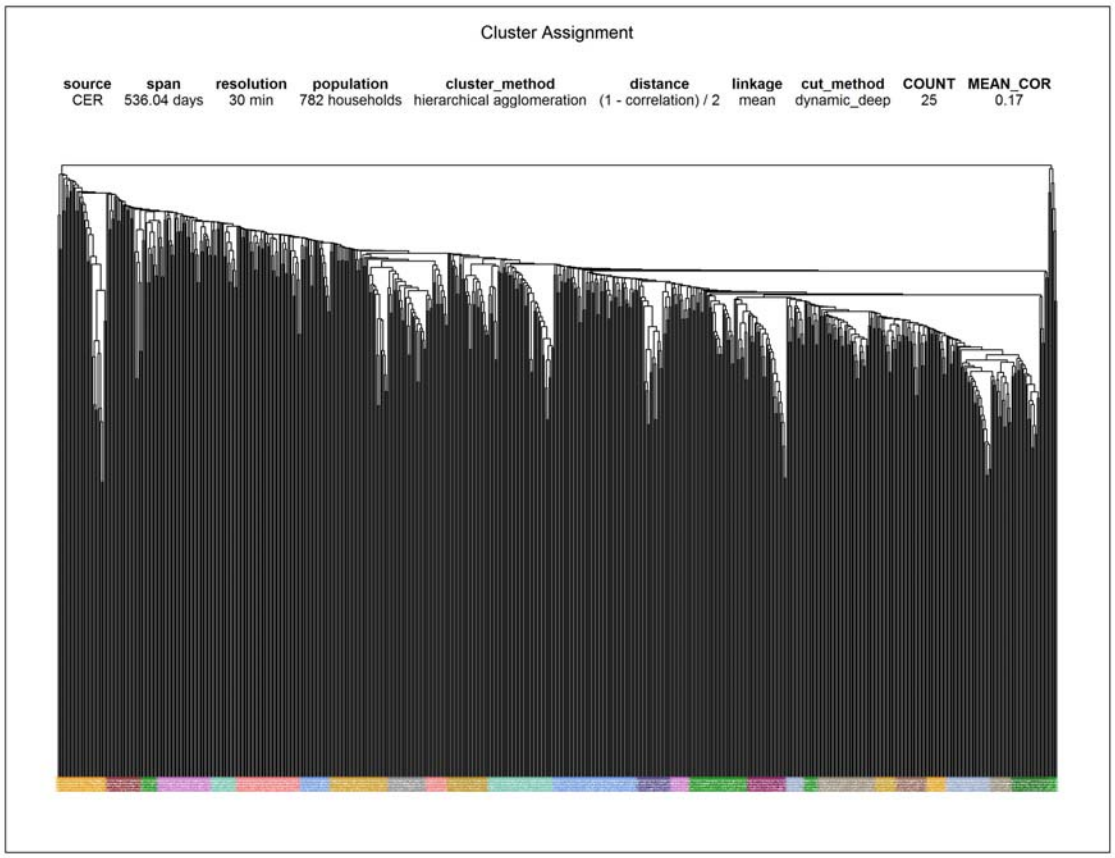

```
4 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  782 (distance mean 0.431, sd 0.038, max 0.672, min 0.239  :::  cor mean 0.138, sd 0.075, max 0.522, min -0.343)
4 clusters of sizes ...
  480 (distance mean 0.452, sd 0.028, max 0.651, min 0.239  :::  cor mean 0.096, sd 0.057, max 0.522, min -0.303)
  205 (distance mean 0.372, sd 0.023, max 0.457, min 0.244  :::  cor mean 0.256, sd 0.047, max 0.511, min 0.085)
   90 (distance mean 0.393, sd 0.026, max 0.506, min 0.242  :::  cor mean 0.214, sd 0.053, max 0.516, min -0.012)
    7 (distance mean 0.474, sd 0.028, max 0.505, min 0.385  :::  cor mean 0.051, sd 0.055, max 0.230, min -0.009)
                                                          weighted mean 0.151
```

Figure 5-3: Group 782 households as 4 clusters, Ireland. *Colors* indicate to which clusters households are assigned.




```
8 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  782 (distance mean 0.431, sd 0.038, max 0.672, min 0.239  :::  cor mean 0.138, sd 0.075, max 0.522, min -0.343)
8 clusters of sizes ...
  205 (distance mean 0.372, sd 0.023, max 0.457, min 0.244  :::  cor mean 0.256, sd 0.047, max 0.511, min 0.085)
  203 (distance mean 0.452, sd 0.023, max 0.567, min 0.301  :::  cor mean 0.095, sd 0.046, max 0.399, min -0.133)
   97 (distance mean 0.395, sd 0.026, max 0.506, min 0.242  :::  cor mean 0.210, sd 0.052, max 0.516, min -0.012)
   85 (distance mean 0.420, sd 0.019, max 0.477, min 0.285  :::  cor mean 0.160, sd 0.038, max 0.429, min 0.045)
   56 (distance mean 0.422, sd 0.021, max 0.493, min 0.338  :::  cor mean 0.156, sd 0.042, max 0.324, min 0.013)
   47 (distance mean 0.473, sd 0.021, max 0.542, min 0.322  :::  cor mean 0.053, sd 0.042, max 0.355, min -0.084)
   44 (distance mean 0.407, sd 0.028, max 0.474, min 0.289  :::  cor mean 0.186, sd 0.055, max 0.421, min 0.051)
   45 (distance mean 0.483, sd 0.043, max 0.602, min 0.239  :::  cor mean 0.033, sd 0.087, max 0.522, min -0.203)
                                                          weighted mean 0.162
```

Figure 5-4: Group 782 households as 8 clusters, Ireland. *Colors* indicate to which clusters households are assigned.




```
25 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  782 (distance mean 0.431, sd 0.038, max 0.672, min 0.239  :::  cor mean 0.138, sd 0.075, max 0.522, min -0.343)
25 clusters of sizes ...
   66 (distance mean 0.421, sd 0.013, max 0.473, min 0.368  :::  cor mean 0.158, sd 0.026, max 0.265, min 0.054)
   51 (distance mean 0.416, sd 0.032, max 0.528, min 0.289  :::  cor mean 0.169, sd 0.063, max 0.421, min -0.056)
   50 (distance mean 0.457, sd 0.017, max 0.527, min 0.358  :::  cor mean 0.086, sd 0.034, max 0.284, min -0.054)
   45 (distance mean 0.431, sd 0.023, max 0.482, min 0.301  :::  cor mean 0.138, sd 0.047, max 0.399, min 0.036)
   45 (distance mean 0.375, sd 0.014, max 0.425, min 0.322  :::  cor mean 0.251, sd 0.029, max 0.356, min 0.150)
   42 (distance mean 0.469, sd 0.018, max 0.520, min 0.400  :::  cor mean 0.062, sd 0.036, max 0.200, min -0.040)
   38 (distance mean 0.473, sd 0.045, max 0.557, min 0.239  :::  cor mean 0.054, sd 0.090, max 0.522, min -0.115)
   35 (distance mean 0.338, sd 0.025, max 0.389, min 0.244  :::  cor mean 0.324, sd 0.051, max 0.511, min 0.221)
   34 (distance mean 0.396, sd 0.019, max 0.448, min 0.334  :::  cor mean 0.209, sd 0.039, max 0.332, min 0.104)
   32 (distance mean 0.422, sd 0.023, max 0.493, min 0.346  :::  cor mean 0.157, sd 0.045, max 0.308, min 0.013)
   30 (distance mean 0.395, sd 0.019, max 0.454, min 0.320  :::  cor mean 0.210, sd 0.038, max 0.360, min 0.091)
   30 (distance mean 0.357, sd 0.028, max 0.426, min 0.242  :::  cor mean 0.287, sd 0.055, max 0.516, min 0.149)
   29 (distance mean 0.353, sd 0.033, max 0.457, min 0.267  :::  cor mean 0.294, sd 0.067, max 0.467, min 0.085)
   28 (distance mean 0.477, sd 0.020, max 0.515, min 0.322  :::  cor mean 0.047, sd 0.040, max 0.355, min -0.029)
   26 (distance mean 0.394, sd 0.034, max 0.458, min 0.285  :::  cor mean 0.211, sd 0.068, max 0.429, min 0.084)
   24 (distance mean 0.377, sd 0.016, max 0.431, min 0.309  :::  cor mean 0.245, sd 0.033, max 0.383, min 0.138)
   23 (distance mean 0.448, sd 0.019, max 0.487, min 0.376  :::  cor mean 0.105, sd 0.038, max 0.247, min 0.026)
   20 (distance mean 0.460, sd 0.024, max 0.517, min 0.396  :::  cor mean 0.081, sd 0.048, max 0.208, min -0.034)
   17 (distance mean 0.407, sd 0.017, max 0.455, min 0.365  :::  cor mean 0.186, sd 0.034, max 0.269, min 0.090)
   16 (distance mean 0.373, sd 0.015, max 0.407, min 0.328  :::  cor mean 0.254, sd 0.029, max 0.345, min 0.186)
   16 (distance mean 0.330, sd 0.020, max 0.392, min 0.283  :::  cor mean 0.340, sd 0.039, max 0.433, min 0.217)
   15 (distance mean 0.410, sd 0.018, max 0.453, min 0.370  :::  cor mean 0.179, sd 0.036, max 0.259, min 0.094)
   15 (distance mean 0.369, sd 0.013, max 0.392, min 0.326  :::  cor mean 0.262, sd 0.025, max 0.349, min 0.216)
   14 (distance mean 0.391, sd 0.019, max 0.435, min 0.350  :::  cor mean 0.218, sd 0.038, max 0.301, min 0.130)
   41 (distance mean 0.454, sd 0.044, max 0.580, min 0.322  :::  cor mean 0.092, sd 0.087, max 0.357, min -0.159)
                                                          weighted mean 0.173
```

Figure 5-5: Group 782 households as 25 clusters, Ireland. *Colors* indicate to which clusters households are assigned.

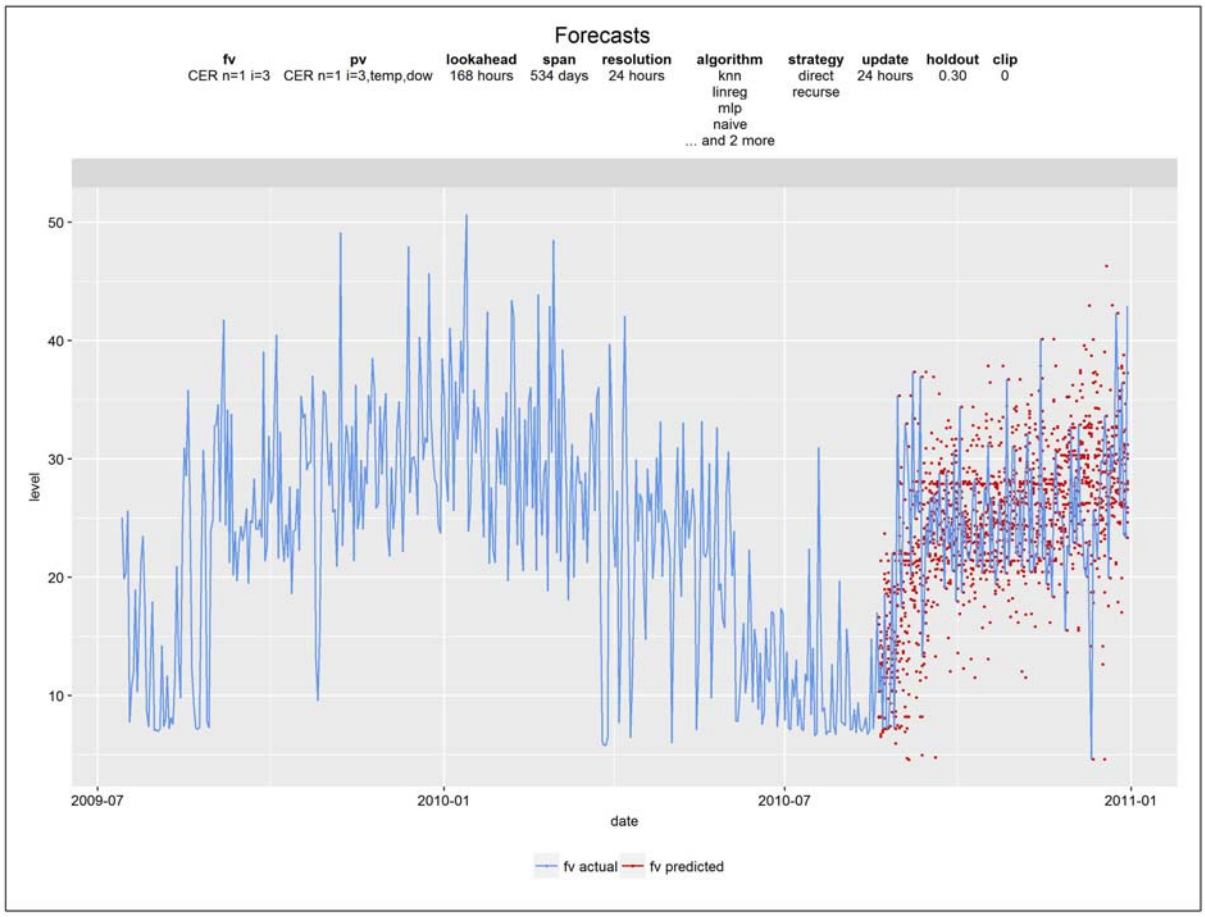


Figure 5-6: Effect of sampling – forecasts, Ireland, week-ahead forecasts. 12 techniques. *Blue* is electricity usage/demand series for a specific sample of size 1 out of 782. *Red* is overlay of 12 forecasts of the specific sample.

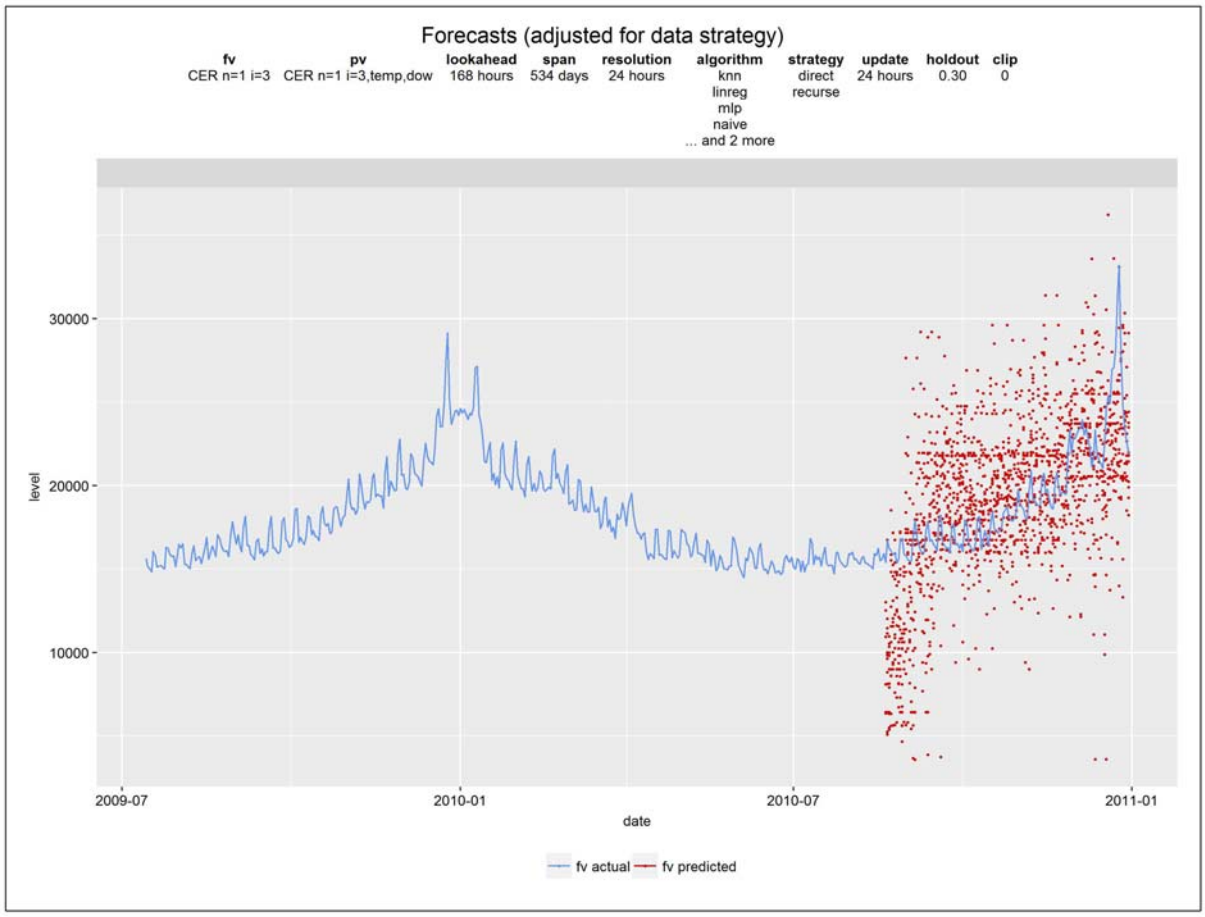


Figure 5-7: Effect of sampling – forecasts, Ireland, week-ahead forecast. *Blue* is electricity usage/demand series for the population. *Red* is overlay of 12 forecasts of a specific sample of size 1 out of 782 re-scaled to the population size.

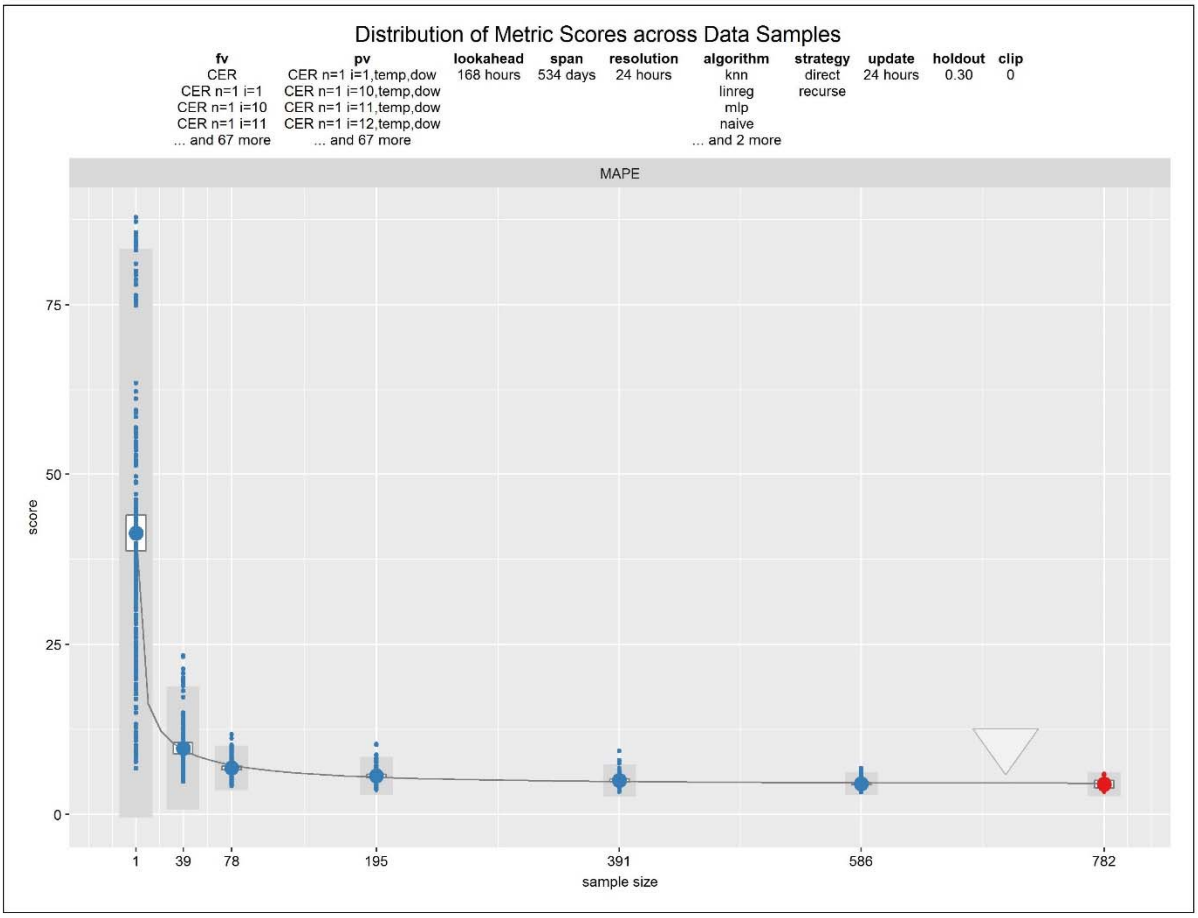


Figure 5-8: Effect of sampling – distribution of metric scores vs. sample size, Ireland, week-ahead forecasts. 12 techniques. 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. Metric is MAPE. *Blue* is metric score for a forecast at a specific sample size. *Large blue* is mean metric score of forecasts at a specific sample size. *Red* is metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

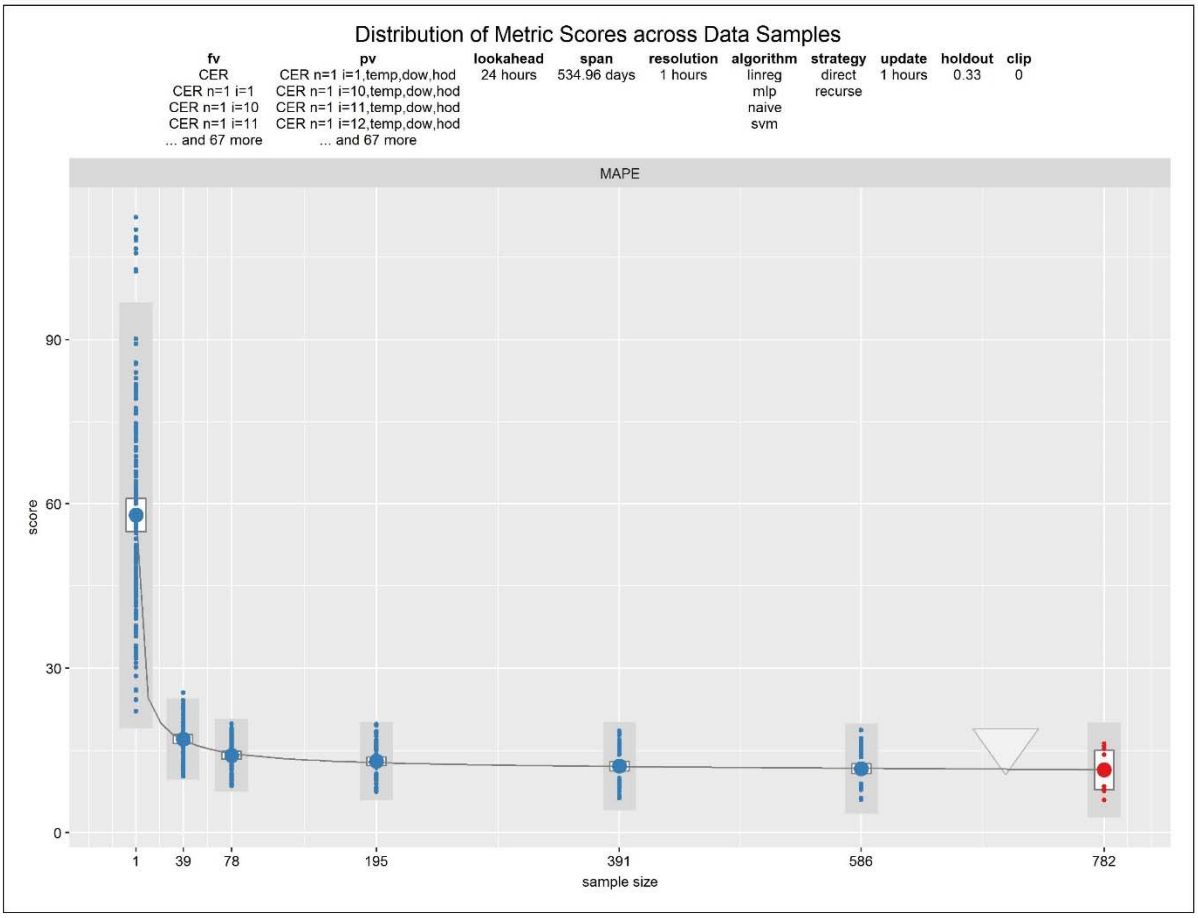


Figure 5-9: Effect of sampling – distribution of metric scores vs. sample size, Ireland, day-ahead forecasts, effect of sampling. 8 techniques. 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. Metric is MAPE. *Blue* is metric score for a specific forecast at a specific sample size. *Large blue* indicates mean metric score of forecasts at a specific sample size. *Red* indicates metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

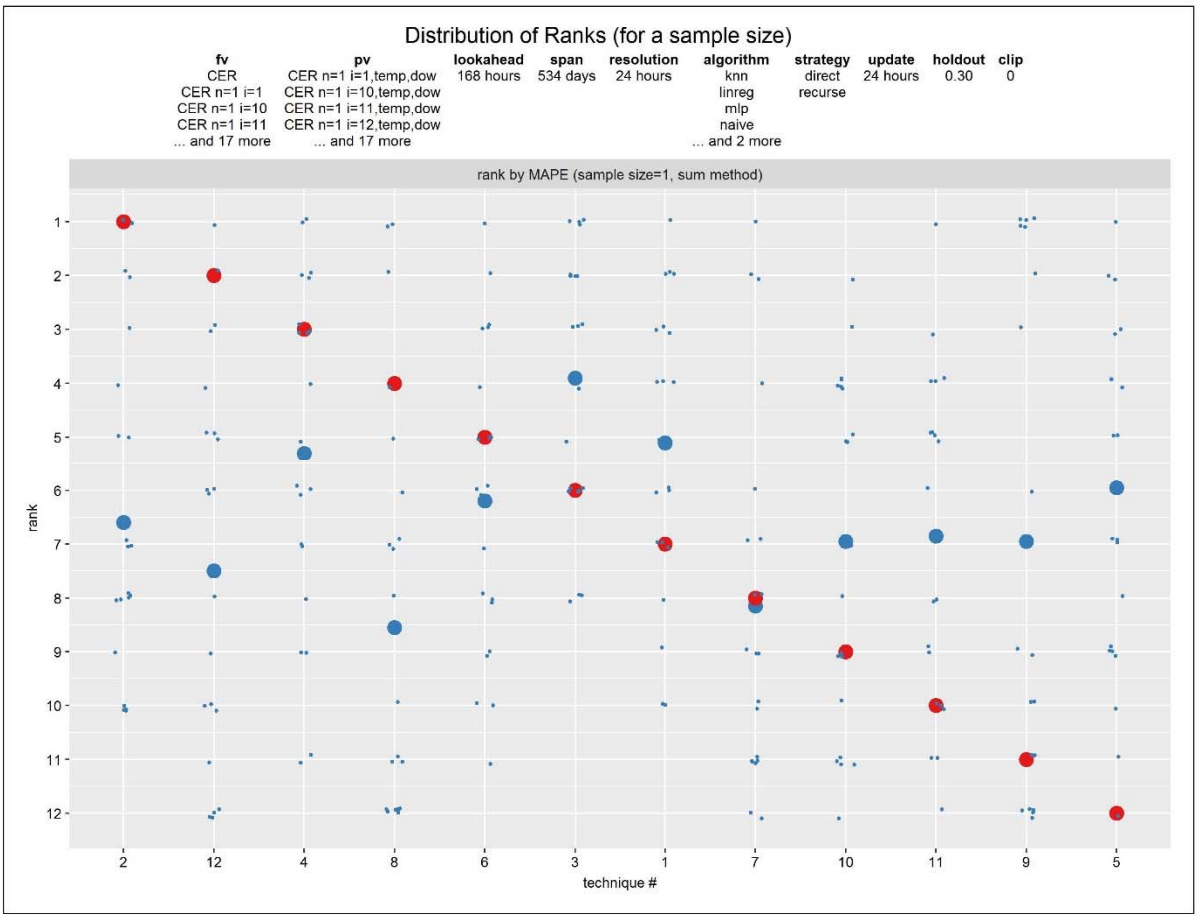


Figure 5-10: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 1, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 20 samples of size 1 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 20 samples. *Large red* is technique rank for a specific technique applied to full population.

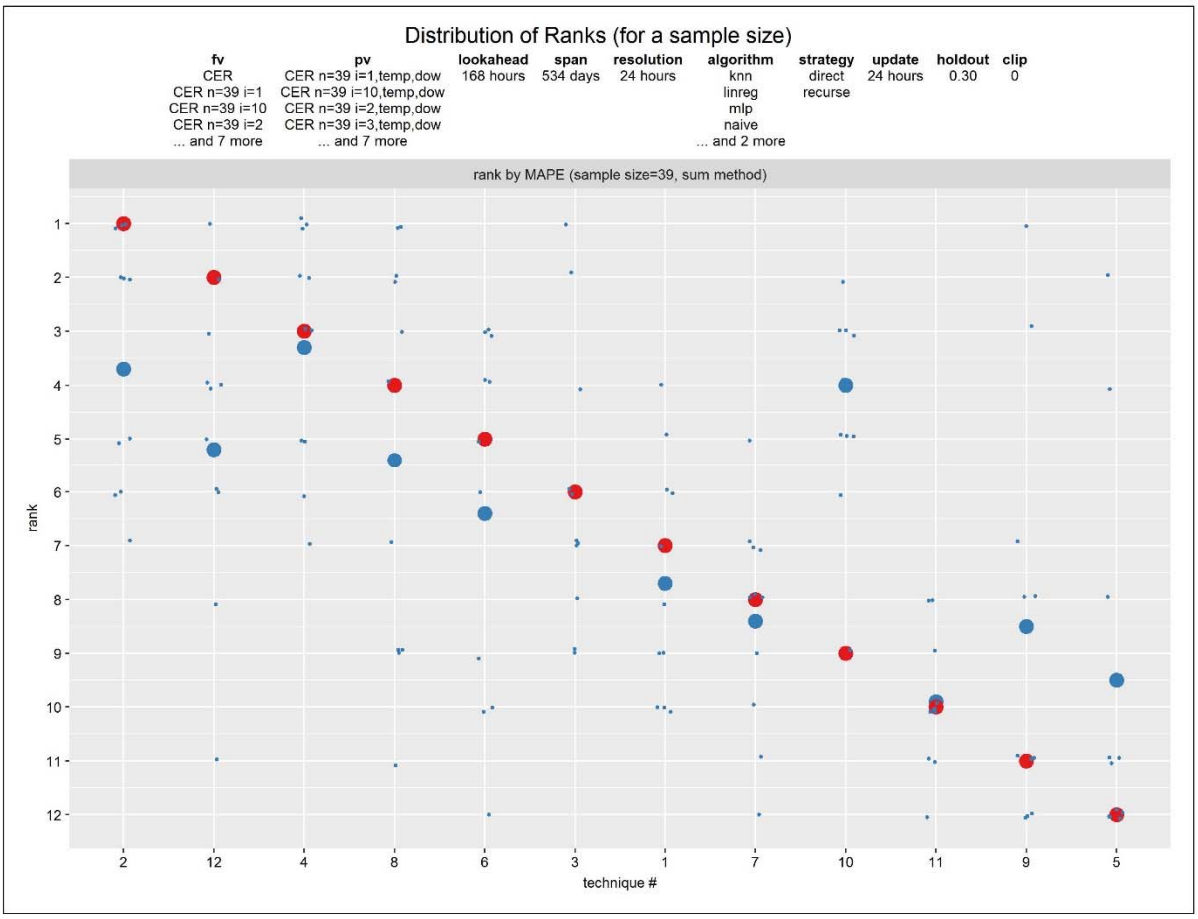


Figure 5-11: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 39, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 39 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

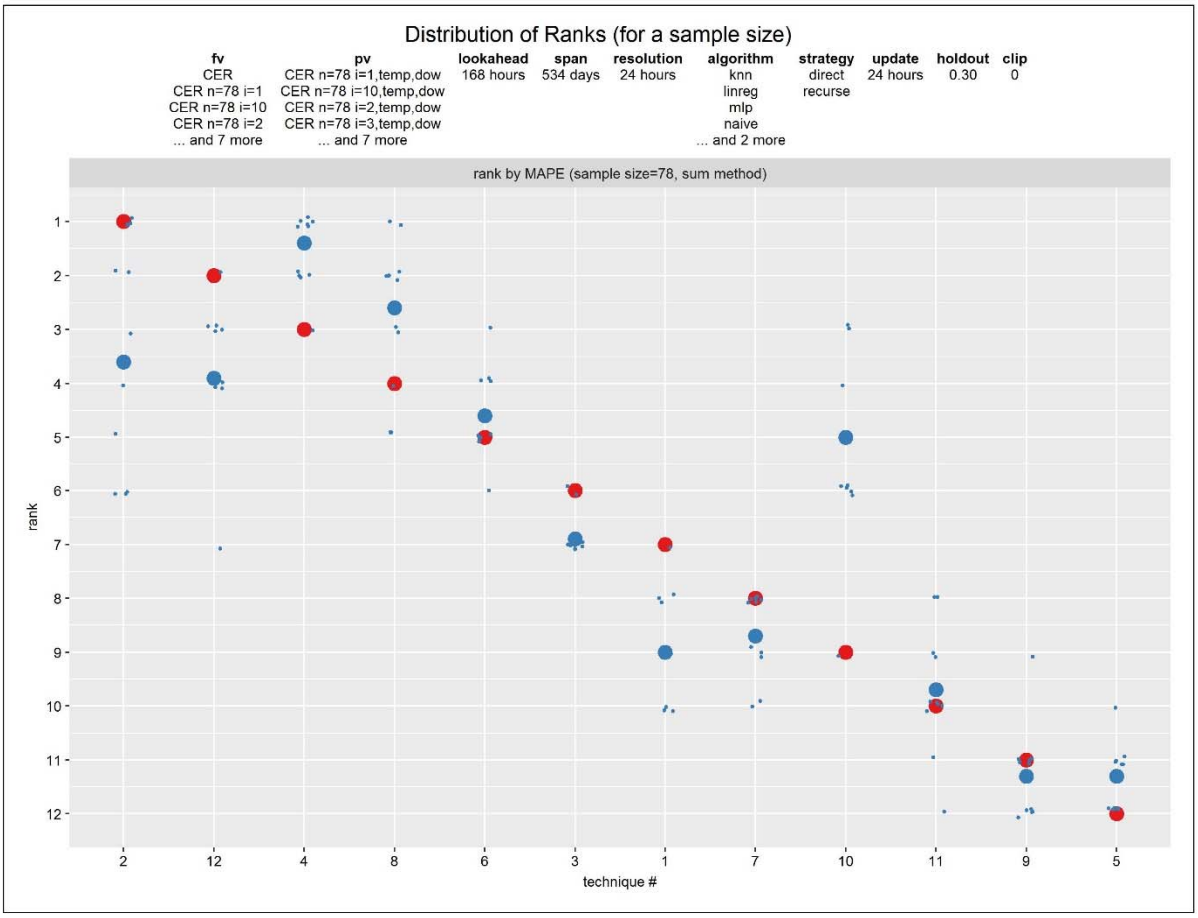


Figure 5-12: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 78, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 78 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

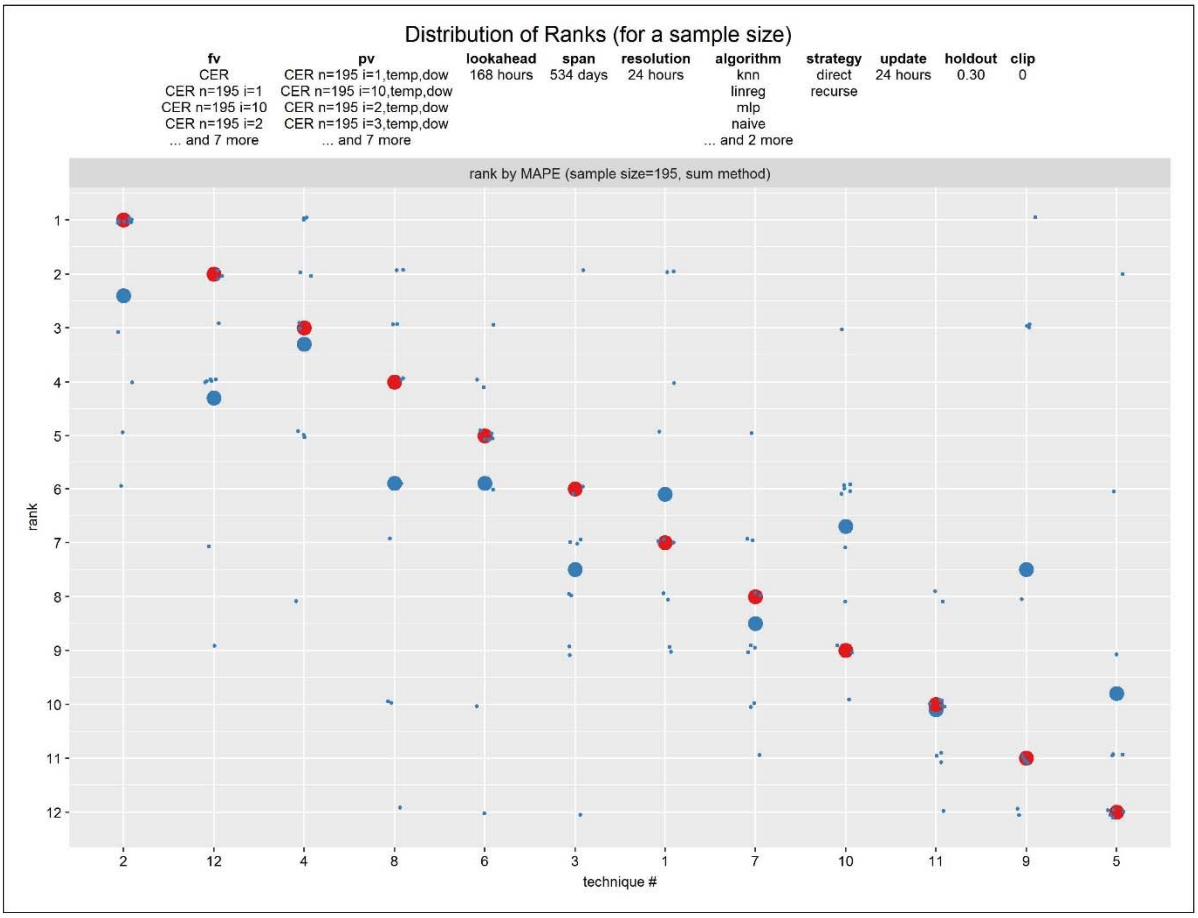


Figure 5-13: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 195, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 195 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

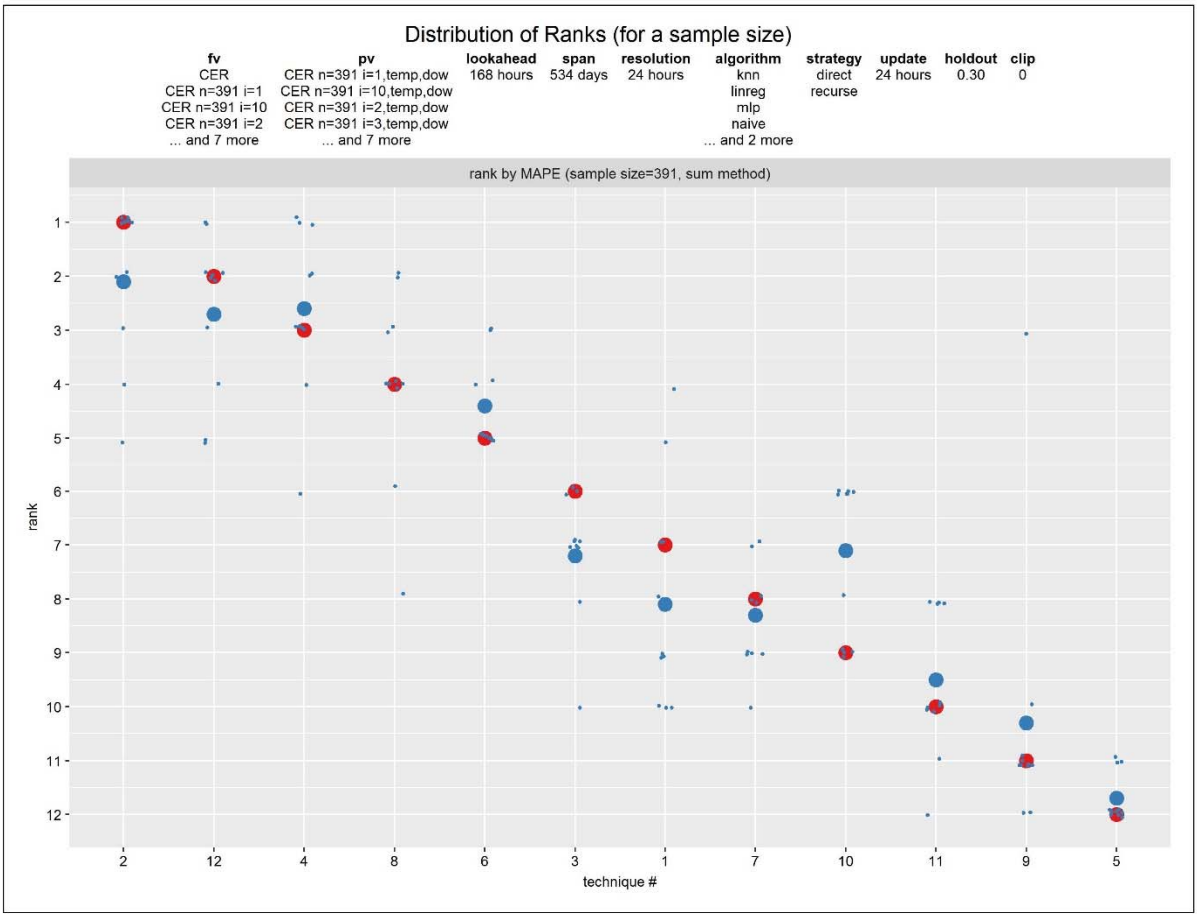


Figure 5-14: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 391, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 1391 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

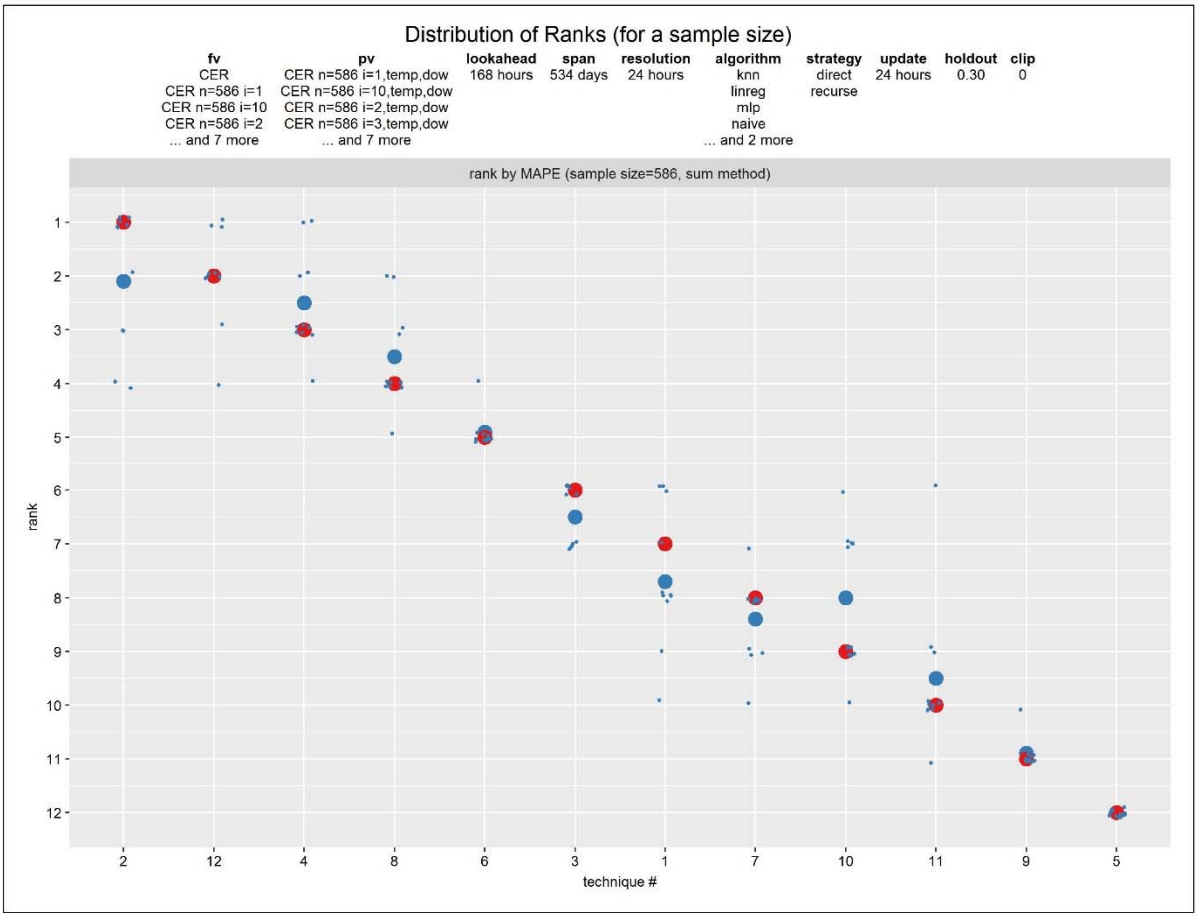


Figure 5-15: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 586, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 586 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

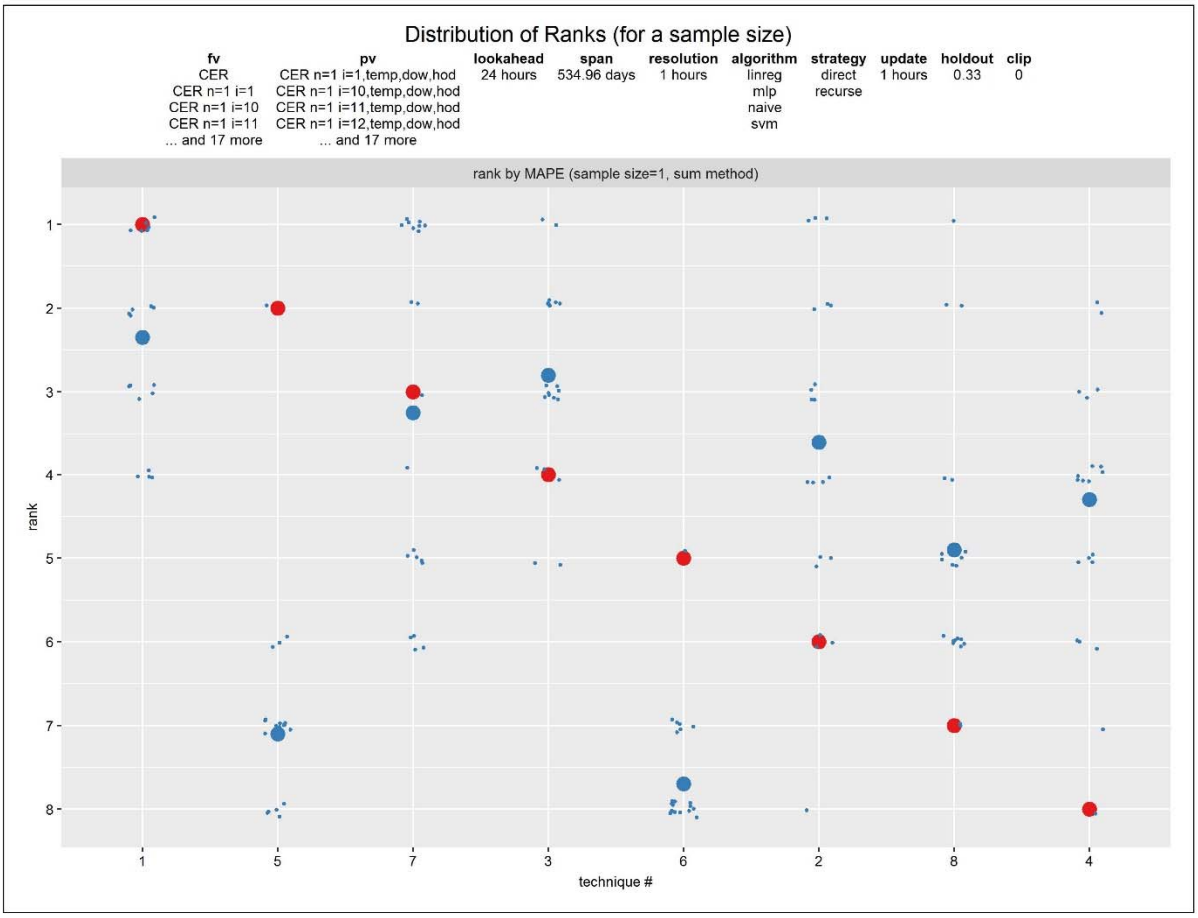


Figure 5-16: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 1, Ireland, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 20 samples of size 1 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 20 samples. *Large red* is technique rank for a specific technique applied to full population.

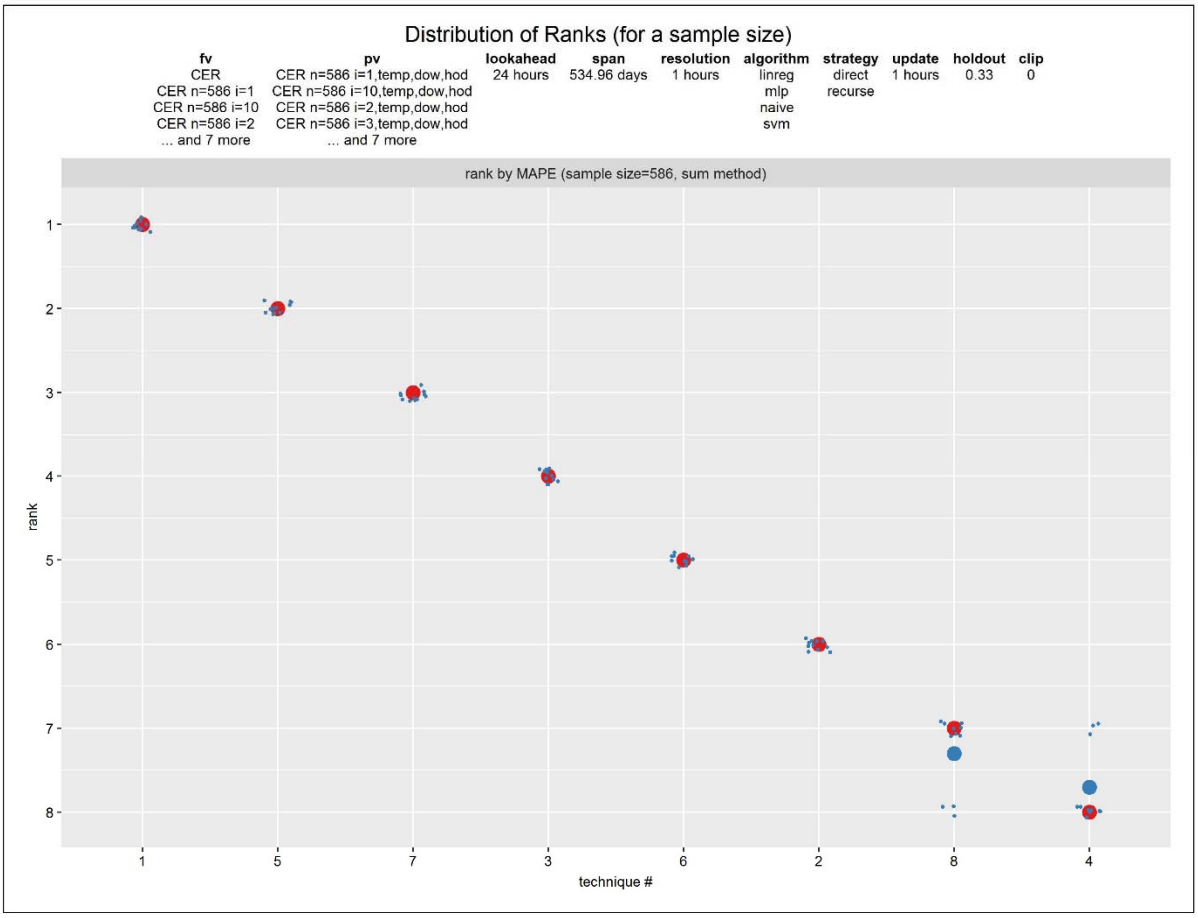


Figure 5-17: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 586, Ireland, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 586 out of 782. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

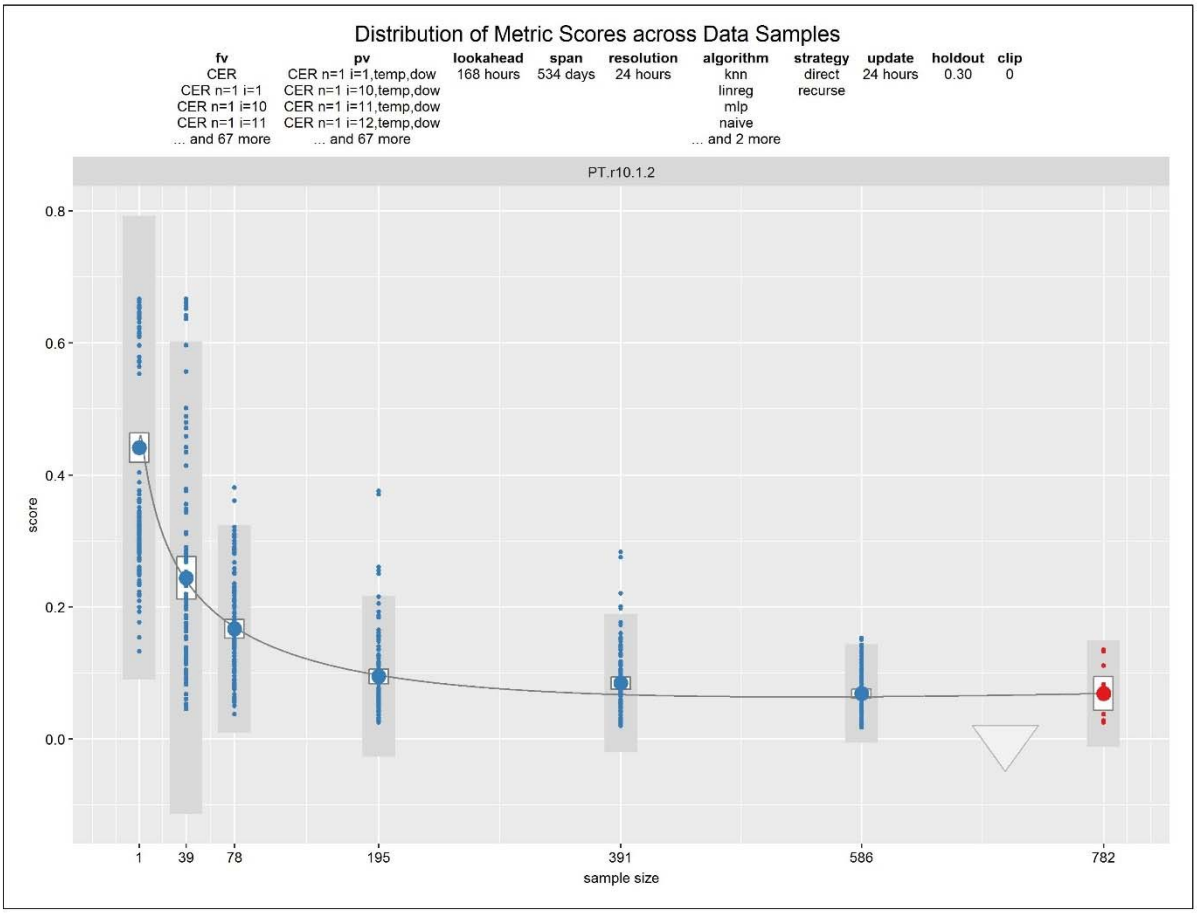


Figure 5-18: Effect of sampling – distribution of metric scores vs. sample size, Ireland, week-ahead forecasts. 12 techniques. 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. Metric is PT.r10.1.2 penalty function, defined as the fraction of time that error exceeds threshold limits, assuming a 10% reserve above forecast, and assuming under-forecasts are twice as costly as over-forecasts. *Blue* is metric score for a forecast at a specific sample size. *Large blue* is mean metric score of forecasts at a specific sample size. *Red* is metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

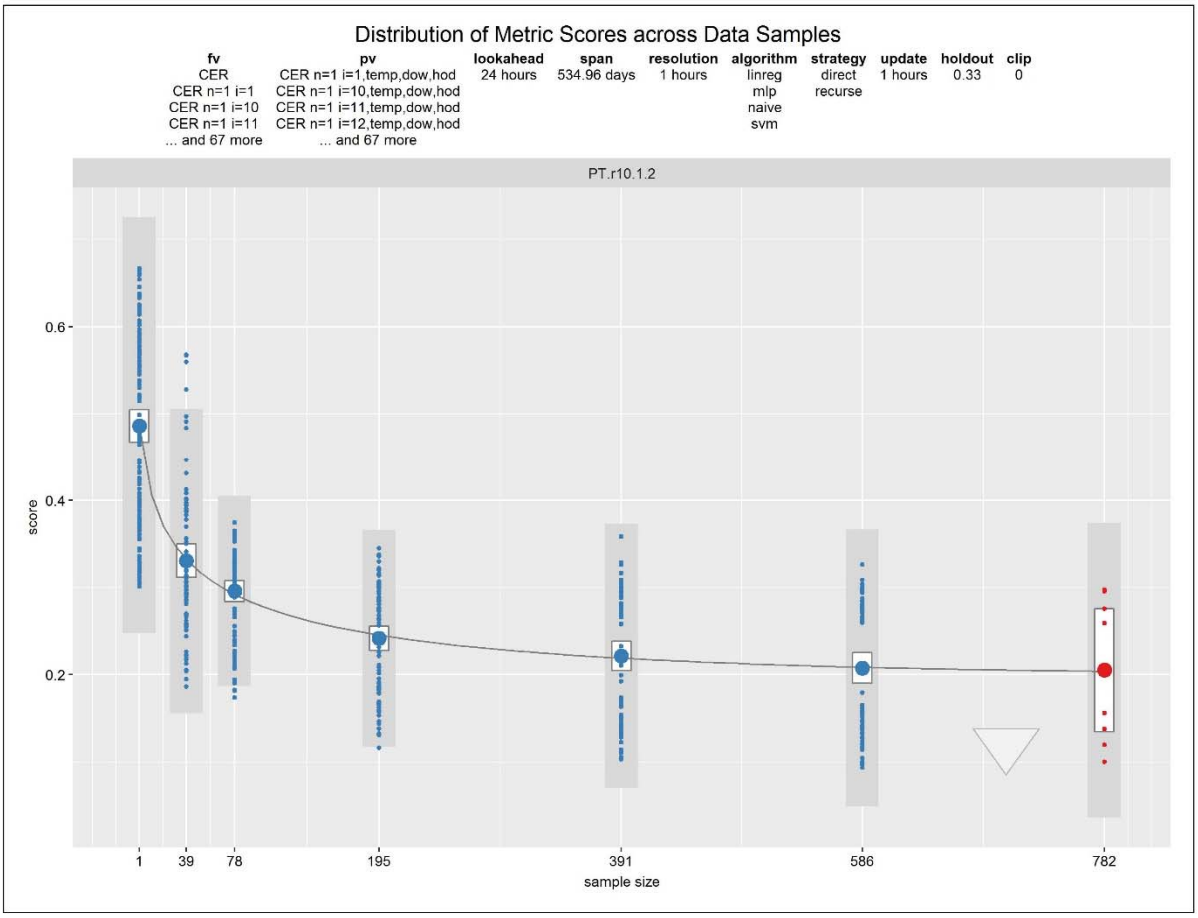


Figure 5-19: Effect of sampling – distribution of metric scores vs. sample size, Ireland, day-ahead forecasts. 8 techniques. 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. Metric is PT.r10.1.2 penalty function, defined as the fraction of time that error exceeds threshold limits, assuming a 10% reserve above forecast, and assuming under-forecasts are twice as costly as over-forecasts. *Blue* is metric score for a forecast at a specific sample size. *Large blue* is mean metric score of forecasts at a specific sample size. *Red* is metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

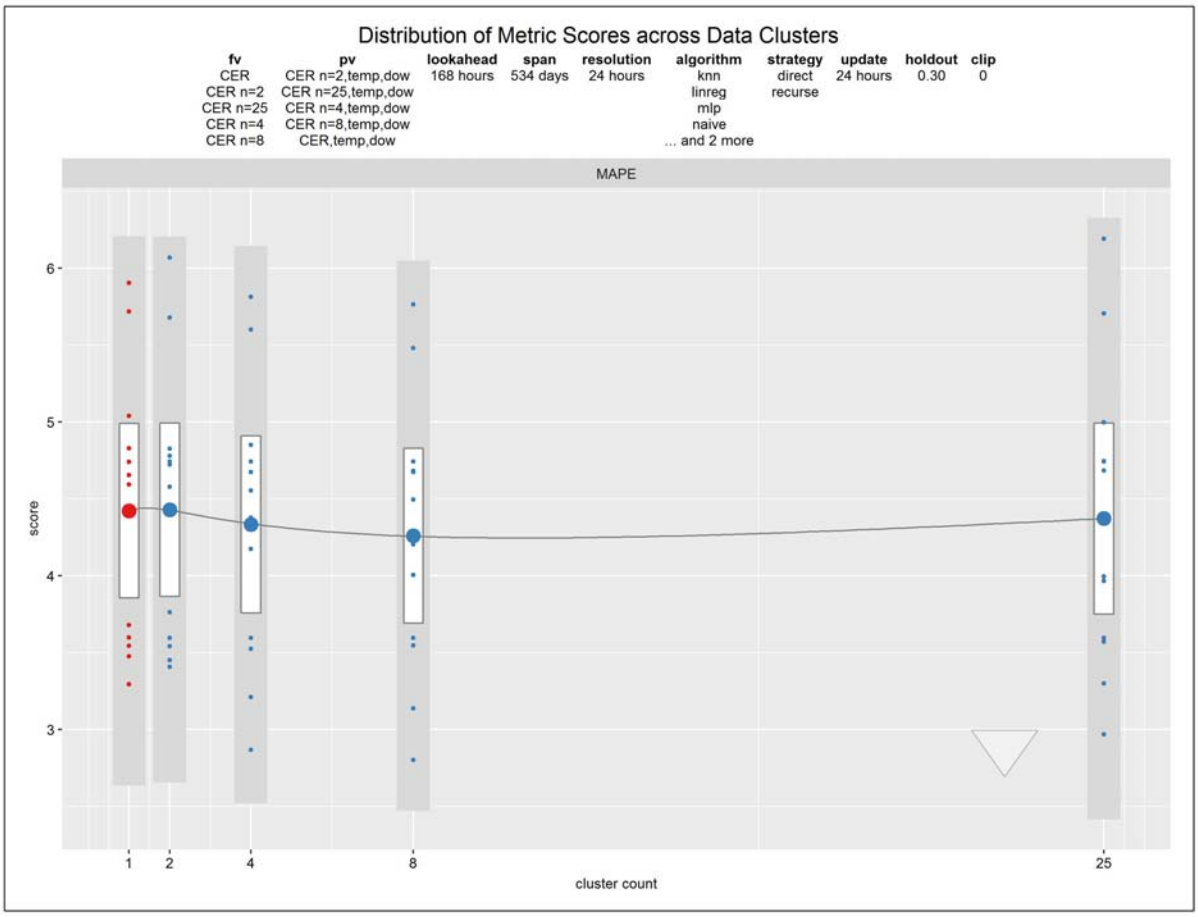


Figure 5-20: Effect of clustering – distribution of metric scores vs. number of clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is metric score for a forecast at a specific number of clusters. *Large blue* is mean metric score of forecasts at a specific number of clusters. *Red* is metric score of a forecast at no clustering (full population). *Large red* is mean metric score of forecasts at no clustering (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific number of clusters. *White* is 95% confidence interval of mean metric score of forecasts at a specific number of clusters.

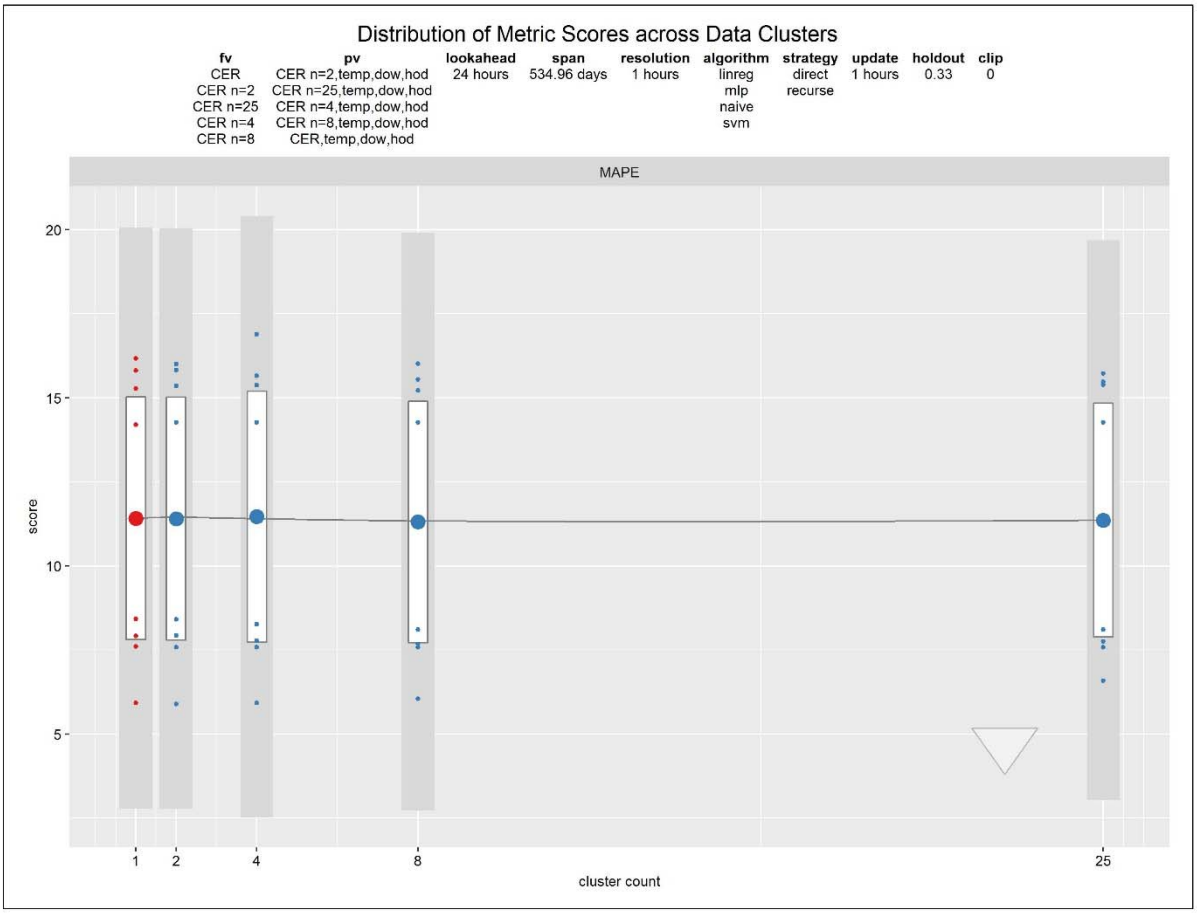


Figure 5-21: Effect of clustering – distribution of metric scores vs. number of clusters, Ireland, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is metric score for a forecast at a specific number of clusters. *Large blue* is mean metric score of forecasts at a specific number of clusters. *Red* is metric score of a forecast at no clustering (full population). *Large red* is mean metric score of forecasts at no clustering (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific number of clusters. *White* is 95% confidence interval of mean metric score of forecasts at a specific number of clusters.

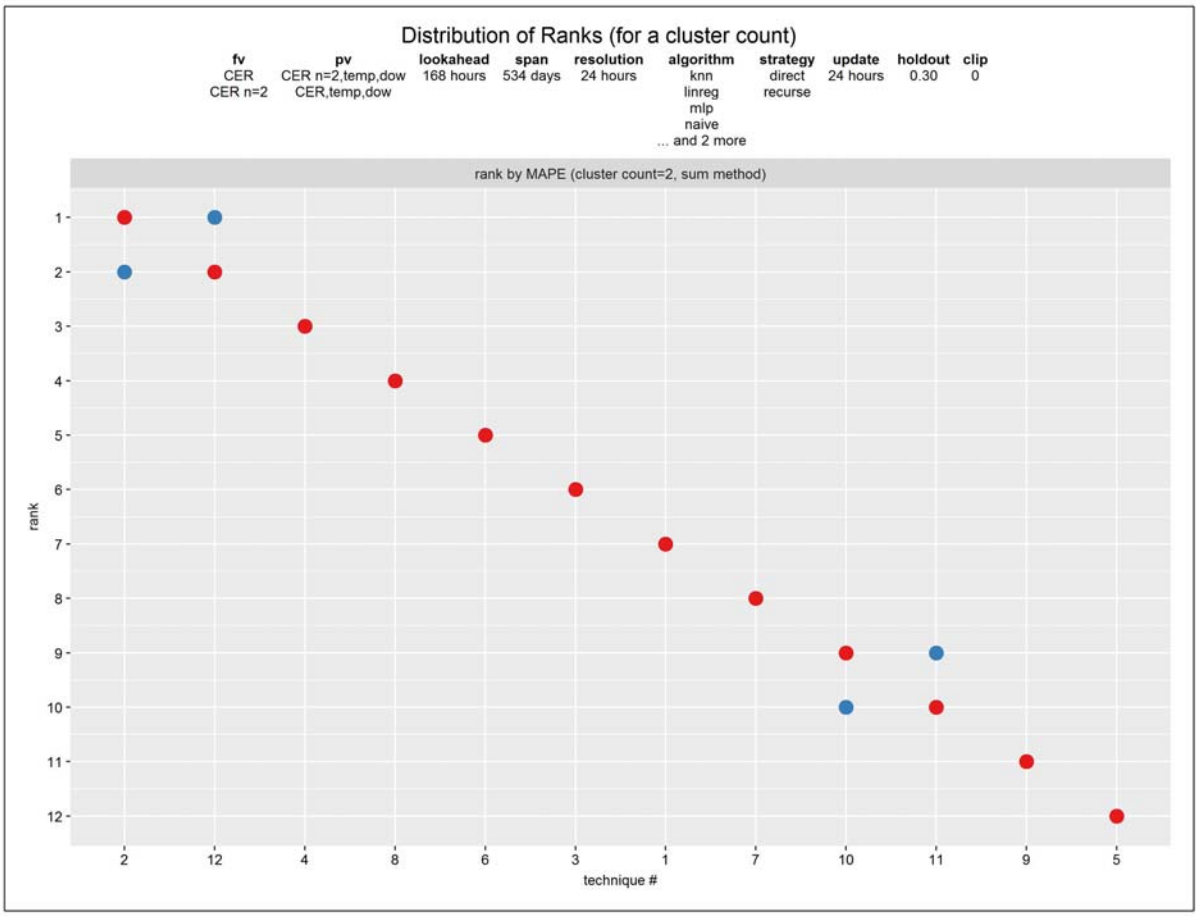


Figure 5-22: Effect of clustering – technique rank vs. technique at 2 clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 2 clusters. *Red* is technique rank for a specific technique applied without clustering.

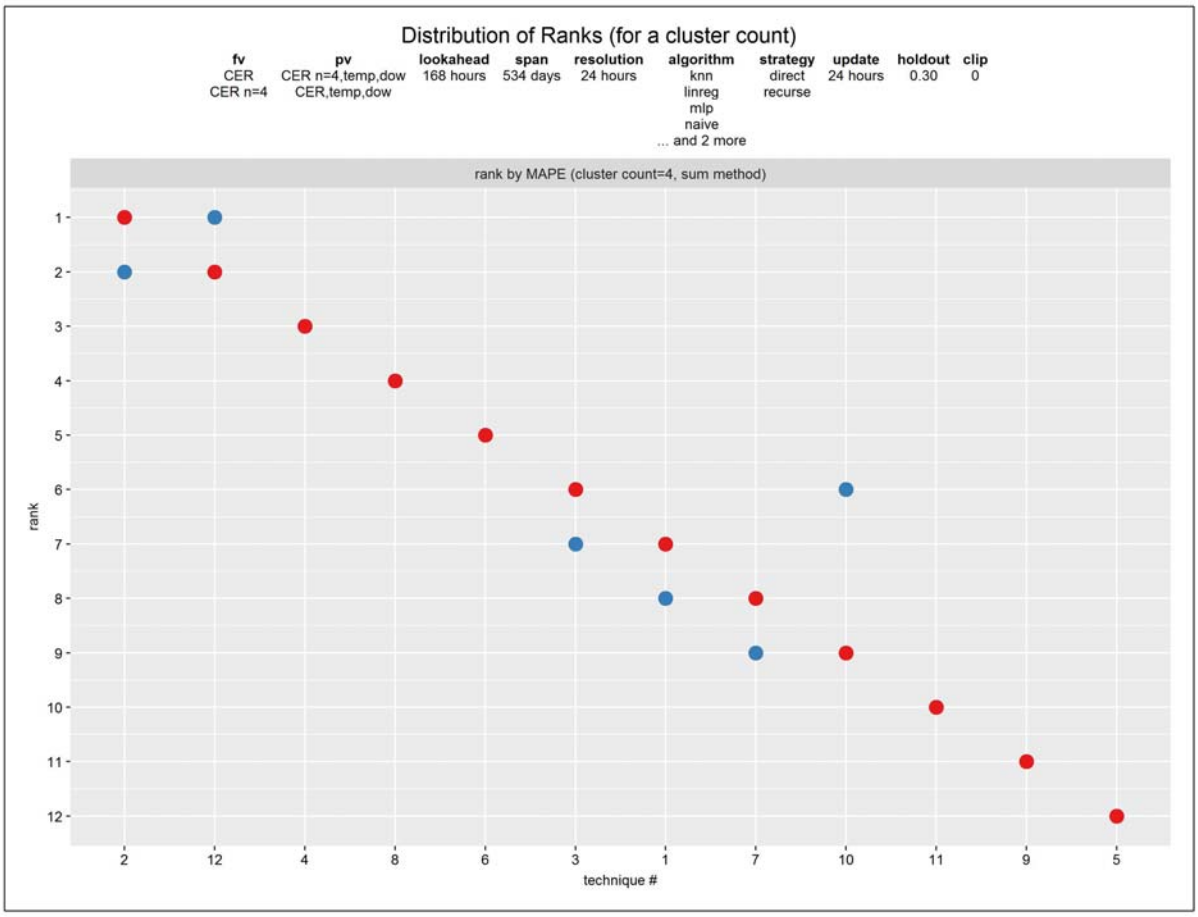


Figure 5-23: Effect of clustering – technique rank vs. technique at 4 clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 4 clusters. *Red* is technique rank for a specific technique applied without clustering.

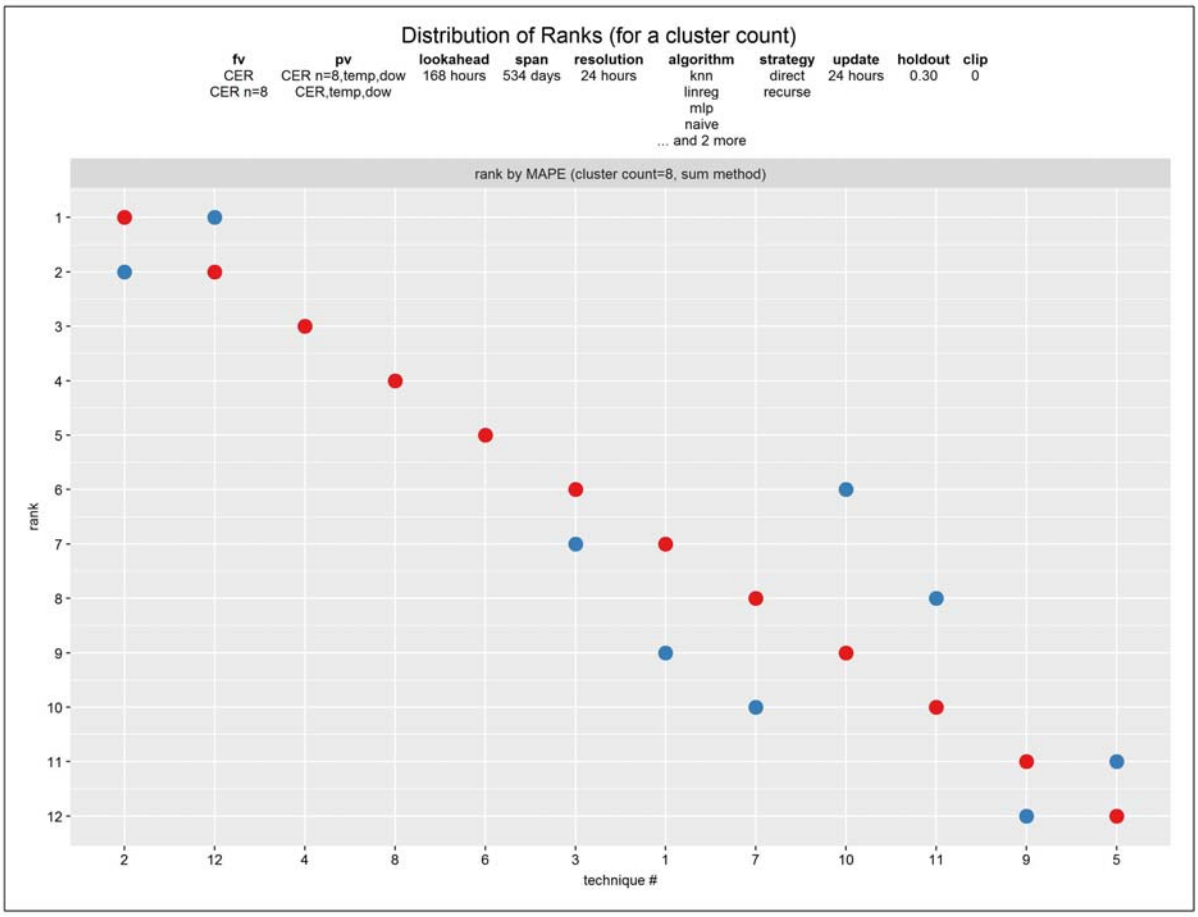


Figure 5-24: Effect of clustering – technique rank vs. technique at 8 clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 8 clusters. *Red* is technique rank for a specific technique applied without clustering.

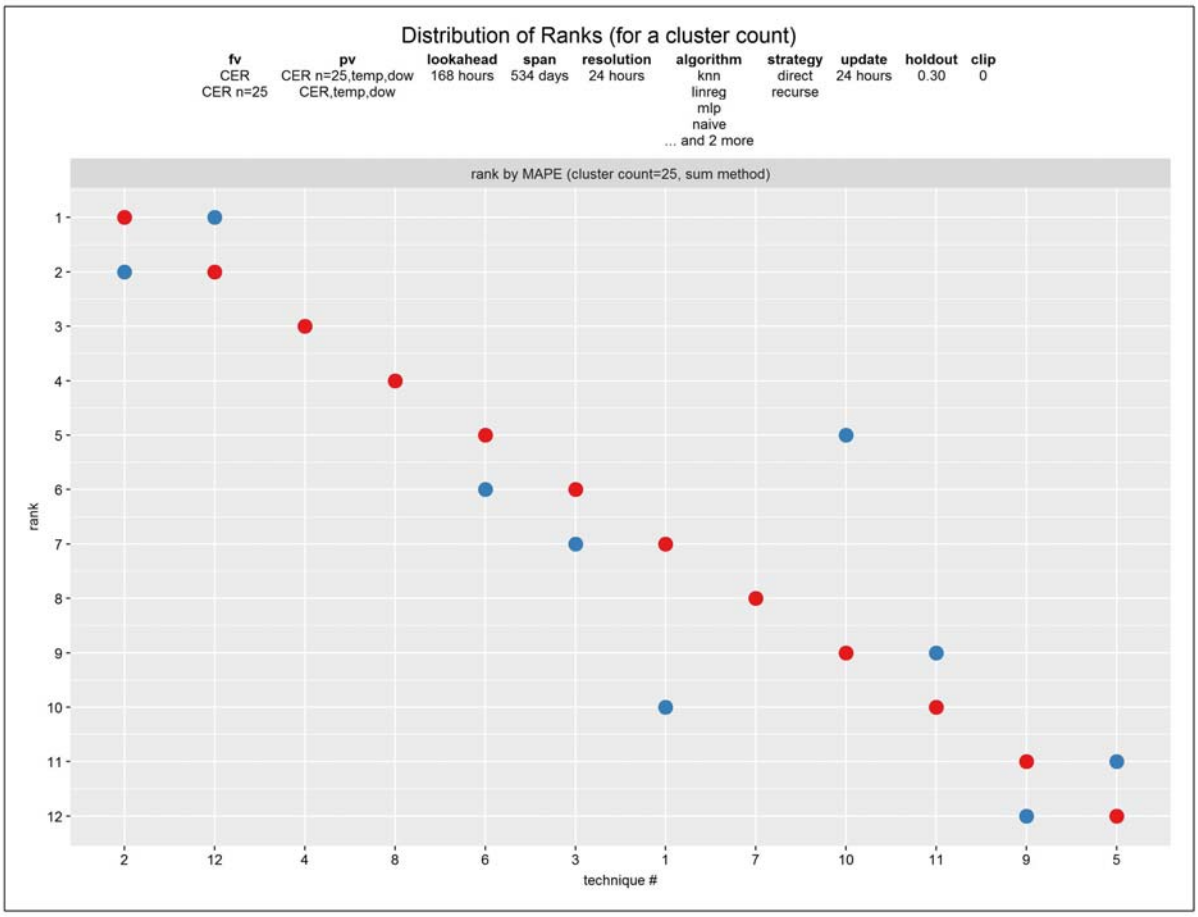


Figure 5-25: Effect of clustering – technique rank vs. technique at 25 clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 25 clusters. *Red* is technique rank for a specific technique applied without clustering.

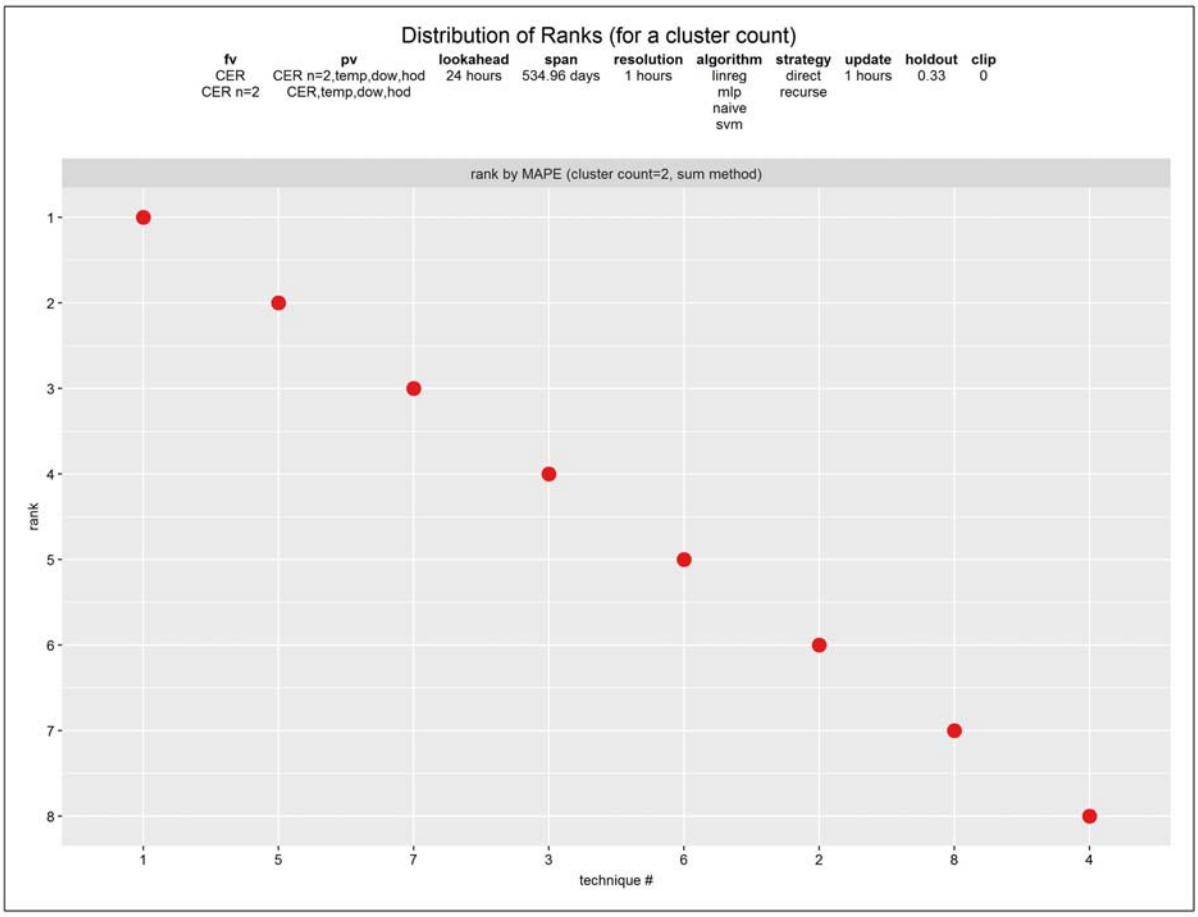


Figure 5-26: Effect of clustering – technique rank vs. technique at 2 clusters, Ireland, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 2 clusters (hidden by red). *Red* is technique rank for a specific technique applied without clustering.

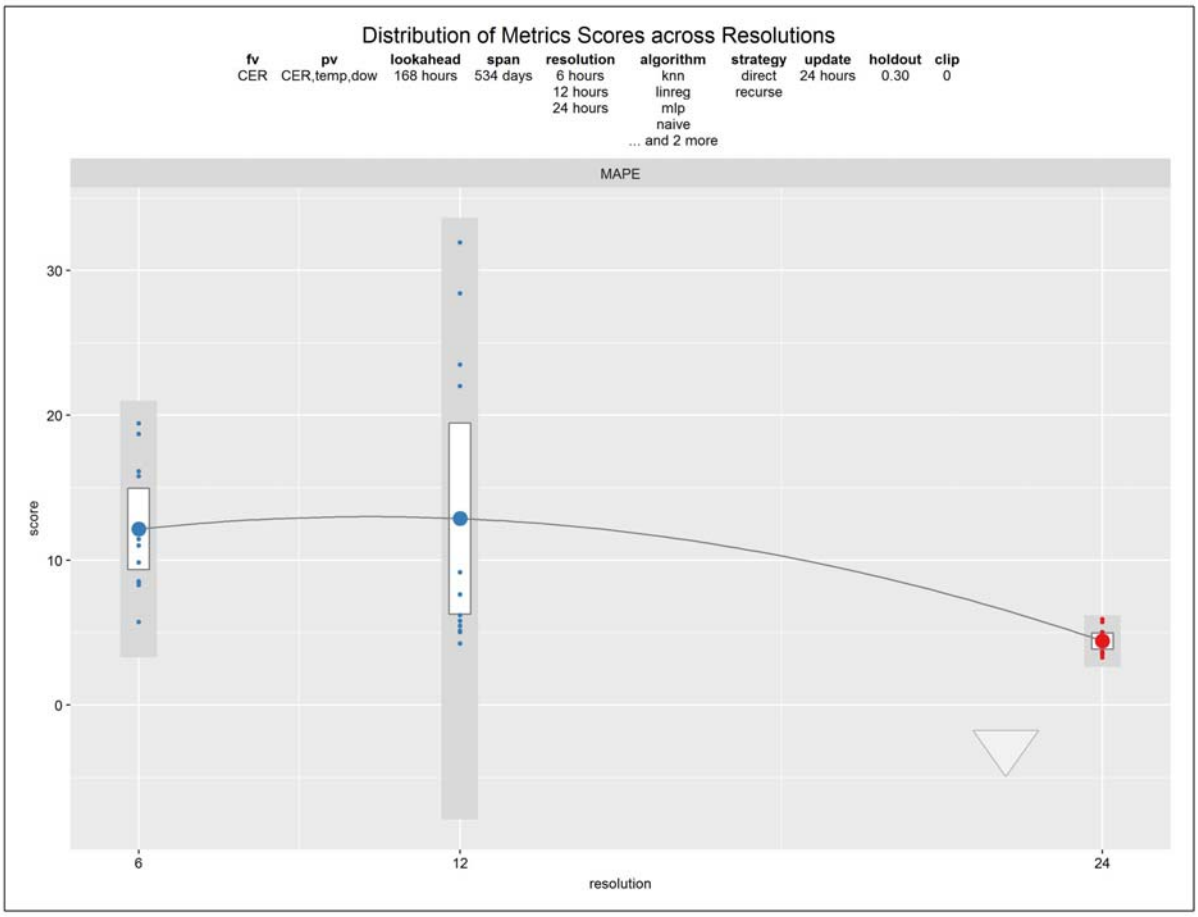


Figure 5-27: Effect of temporal magnification – distribution of metric scores vs. time step size, Ireland, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. *Blue* is metric score for a forecast at a specific time step size. *Large blue* is mean metric score of forecasts at a specific time step size. *Red* is metric score of a forecast at baseline time step size. *Large red* is mean metric score of forecasts at baseline time step size. *Gray* is 2 standard deviations from mean metric score of forecasts at a specific time step size. *White* is 95% confidence interval of mean metric score of forecasts at a specific time step size.

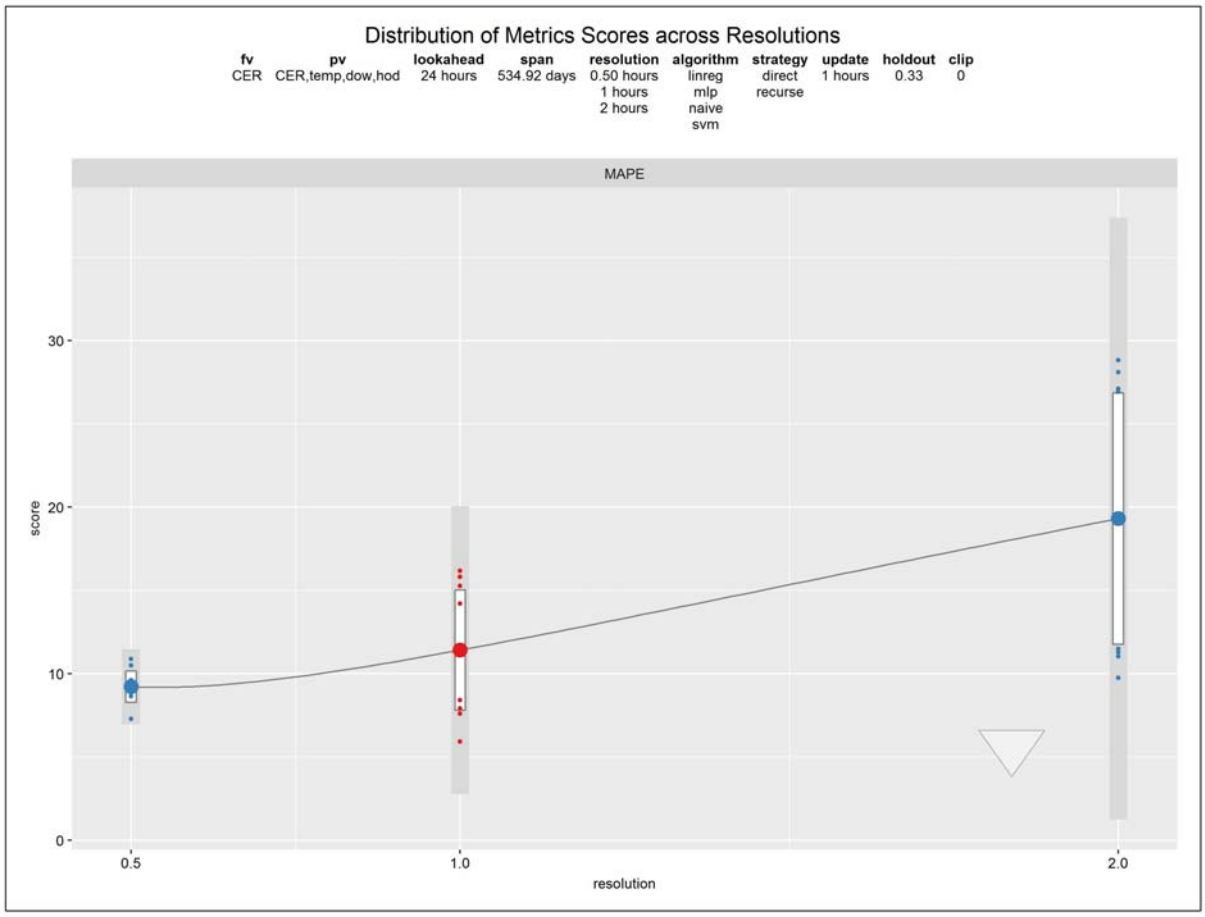


Figure 5-28: Effect of temporal magnification – distribution of metric scores vs. time step size, Ireland, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. *Blue* is metric score for a forecast at a specific time step size. *Large blue* is mean metric score of forecasts at a specific time step size. *Red* is metric score of a forecast at baseline time step size. *Large red* is mean metric score of forecasts at baseline time step size. *Gray* is 2 standard deviations from mean metric score of forecasts at a specific time step size. *White* is 95% confidence interval of mean metric score of forecasts at a specific time step size.

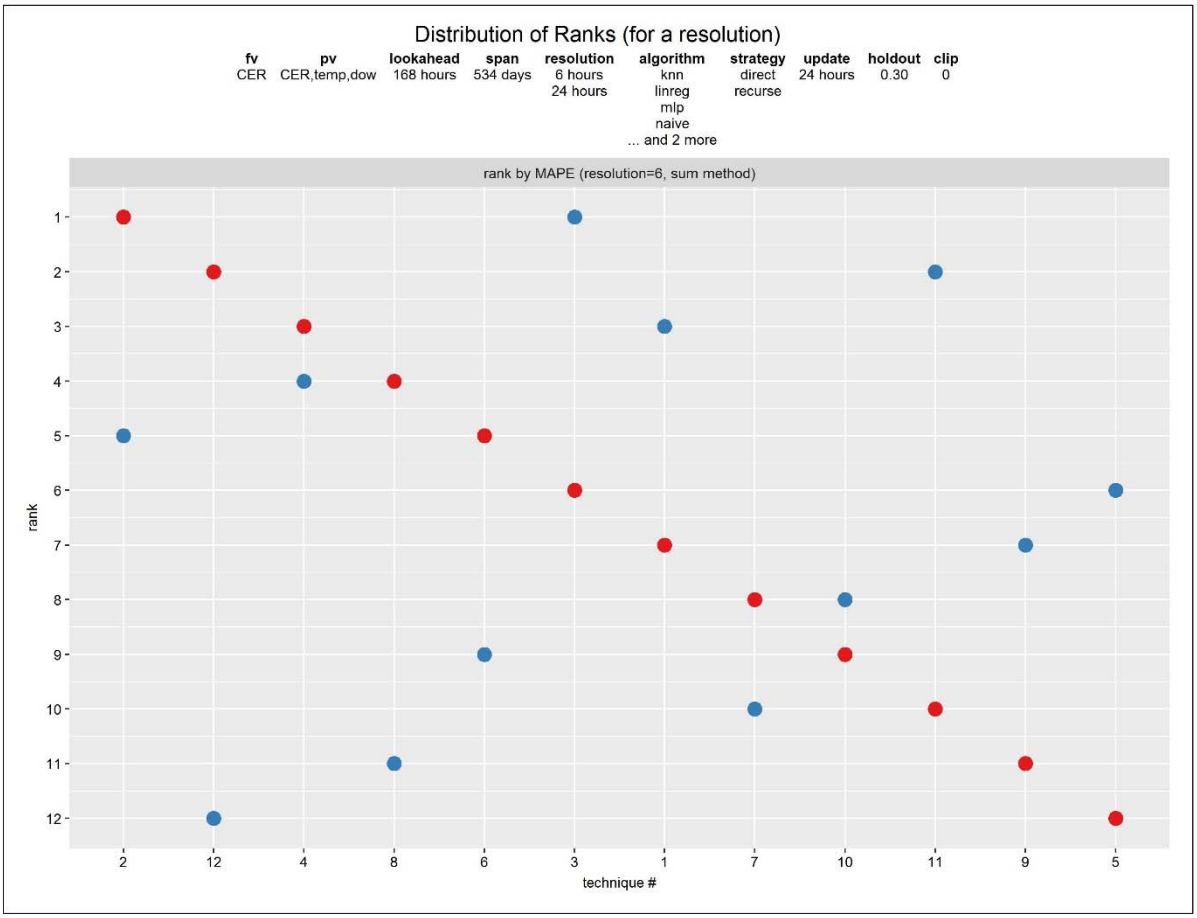


Figure 5-29: Effect of temporal magnification – technique rank vs. technique at short time step size, Ireland, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. *Blue* is technique rank for a specific technique applied at temporal magnification with 6-hour time step size. *Red* is technique rank for a specific technique applied without temporal magnification (baseline 24-hour time step size).

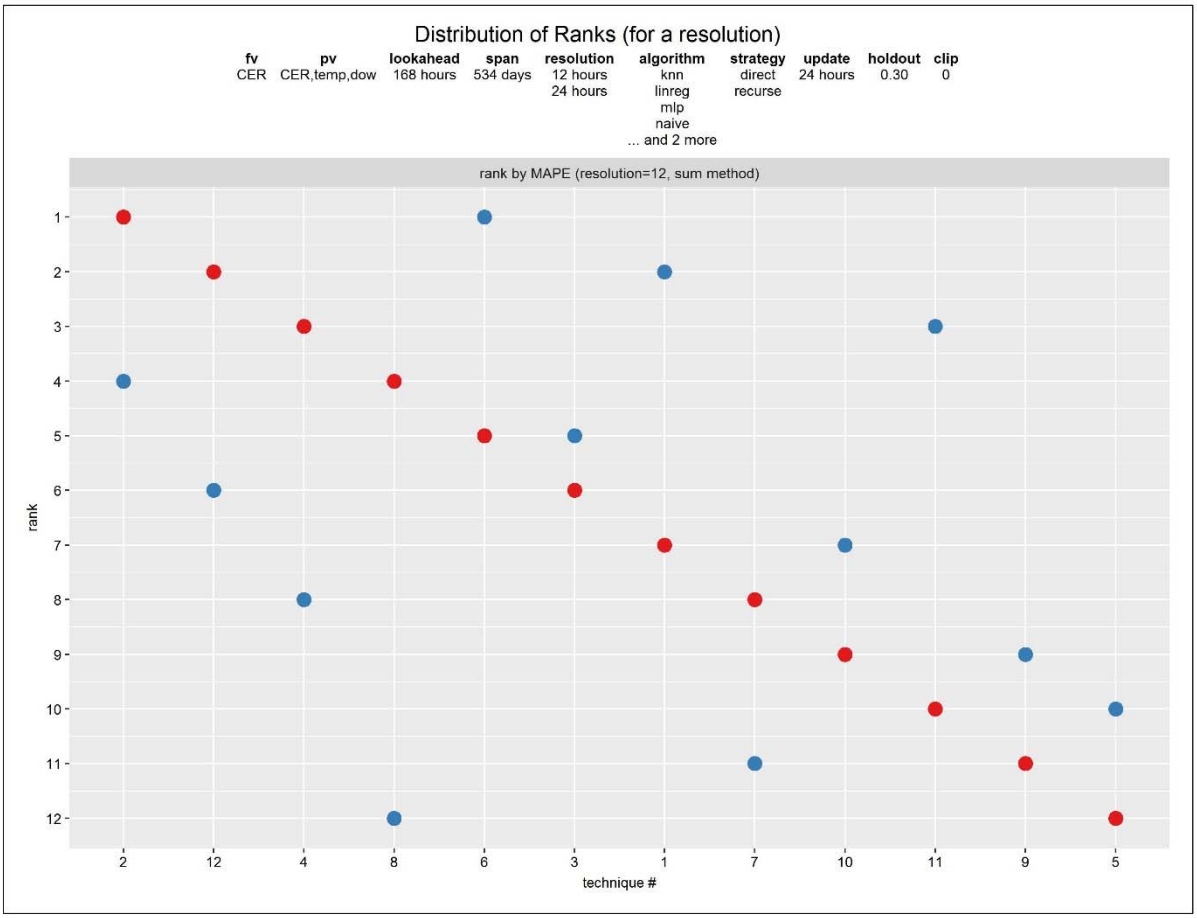


Figure 5-30: Effect of temporal magnification – technique rank vs. technique at medium time step size, Ireland, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. *Blue* is technique rank for a specific technique applied at temporal magnification with 12-hour time step size. *Red* is technique rank for a specific technique applied without temporal magnification (baseline 24-hour time step size).

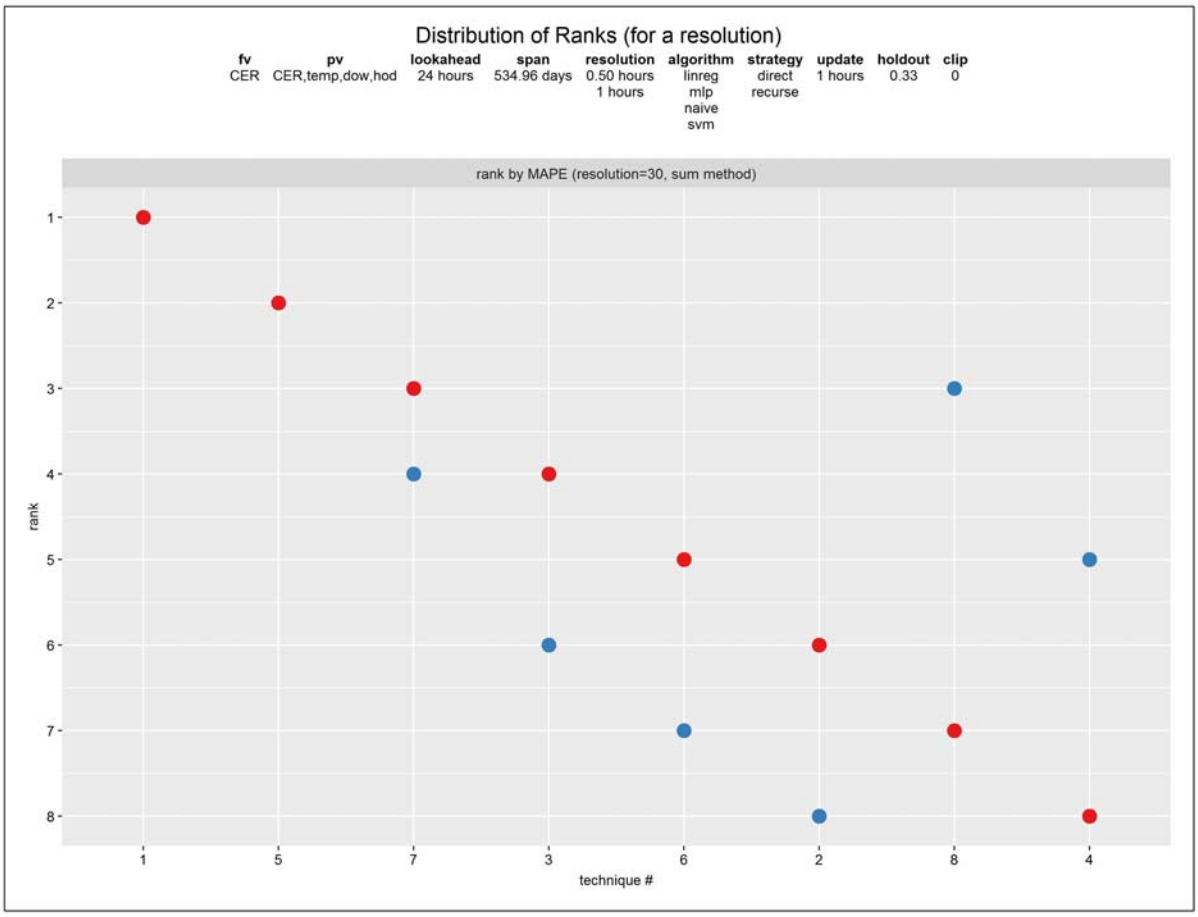


Figure 5-31: Effect of temporal magnification – technique rank vs. technique at short time step size, Ireland, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. *Blue* is technique rank for a specific technique applied at temporal magnification with 30-minute time step size. *Red* is technique rank for a specific technique applied without temporal magnification (baseline 1-hour time step size).

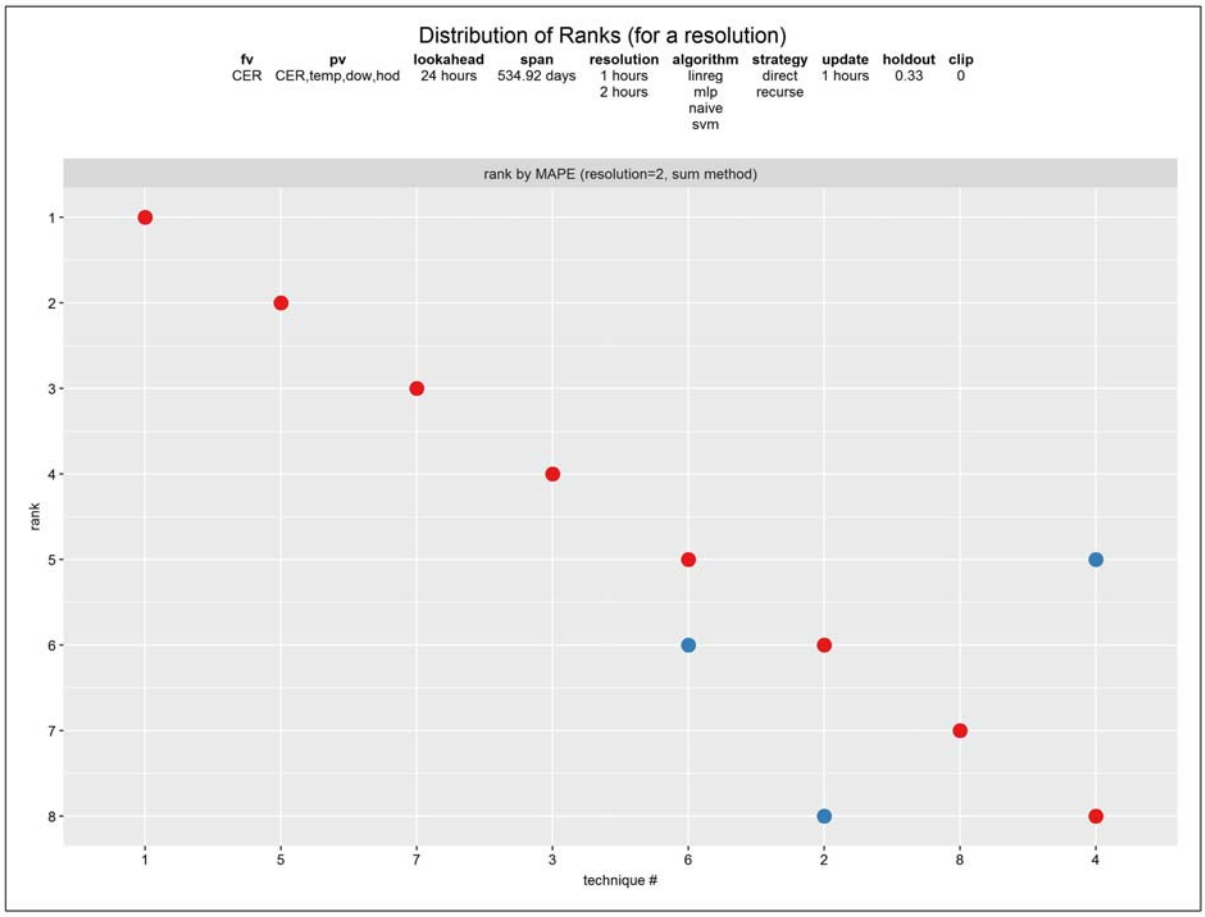


Figure 5-32: Effect of temporal magnification – technique rank vs. technique at long time step size, Ireland, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. *Blue* is technique rank for a specific technique applied at temporal magnification with 2-hour time step size. *Red* is technique rank for a specific technique applied without temporal magnification (baseline 1-hour time step size).

6 ROBUSTNESS OF DECISION EFFECTS AND RESIDENTIAL ELECTRICITY DEMAND ESTIMATION

"It ain't what you don't know that gets you in trouble.
It's what you know for sure that just ain't so."
– Mark Twain

6.1 Research Questions

In this chapter, we address the following the research questions:

- How robust are residential electricity demand forecasting techniques to data sources?
- Are the relationships between forecasting process decisions and residential electricity demand forecasting performance universal or location-specific?

6.2 Research Approach

To explore the robustness of our earlier results to variations in data source, we essentially repeat our analysis of week-ahead forecasting, substituting smart electric grid data from the Australian government for the data from Ireland CER – and then compare the two sets of results.

6.3 Scope of Analysis

We scope our week-ahead analysis, as before, to the effects of 12 decisions, and then drill down on the effects of 3 data strategy decisions. Day-ahead analysis is scoped similarly.

6.3.1 Model Instantiations

Our week-ahead model instantiations vary decisions, as before, but substituting a different reference data source. Our day-ahead model instantiations vary decisions similarly.

Objective Decisions	lock-in	Reference data source	Australia SCSG electricity usage
	lock-in	Reference look-ahead	1 week
	5 options	Update cycle	1 day 2 days 3 days 5 days 7 days
	lock-in	Time step size	1 day
Data Strategy Decisions	3 options	Span	6 months 12 months 18 months
Algorithm Decisions	6 options	Algorithm class	k-nearest neighbor linear regression multilayer perceptron naïve support vector regression decision tree
	lock-in	Predictor data sources	Australia SCSG electricity usage + Sydney airport temperature + day of week
	lock-in	Predictor look-backs	(0, 1 week, 2 weeks, 3 weeks) + (0) + (0)
	2 options	Extension rule	direct recurse
Training and Testing Decisions	4 options	Holdout %	20% 30% 40% 50%

	4 options	Clip %	0 1% 10% 20%
Metric and Penalty Decisions	60 options	Metric function or penalty function	any of 30 metrics functions or 30 penalty functions

Sampling

Data Strategy Decisions	6 options	Sample szie	any of 20 1-household samples any of 10 39-household samples any of 10 78-household samples any of 10 195-household samples any of 10 586-household samples full population

Clustering

Data Strategy Decisions	5 options	Cluster grouping	partitioned into 2 clusters partitioned into 4 clusters partitioned into 8 clusters partitioned into 25 clusters not partitioned

Temporal Magnification

Data Strategy Decisions	3 options	Time step size	6 hour 12 hour 1 day

6.3.2 Data Sources

Our techniques all assume the same reference and predictor series, pre-processed as necessary to prepare them for use by the platform.

For our reference series, we use electricity usage time series data from the Australian Government Smart City / Smart Grid Project, assuming that demand exactly equaled usage in the absence of physical constraints or financial disincentives. [128,155] The smart electric grid covered 78,720 households in southeastern Australia, measuring usage at that geographic granularity at half-hour intervals, spanning October 2010 to February 2014. Of these, we restrict our analysis to the 223 households that had no missing data throughout an overlapping 2-year period. To consolidate the 223 time series, we sum across households at each time step, resulting in a 35,088-step time series. Half-hourly values for the aggregated population range from 27 kWh to 251 kWh, with mean 74 kWh.

Our four predictor series are electricity usage, temperature, day of week, and hour of day (for day-ahead only). For electricity usage, we used the same Australia data. For temperature, we used actual temperature time series data, reported at hourly intervals, from the Sydney airport database. For day of week, we used time series data constructed per an integer coding scheme.

Reference series	Electricity usage. From Australia Government Smart City / Smart Grid Project (SCSG). Measured at 30-minute intervals across 223 households in southeastern Australia in 2010-2012.
Predictor series	Electricity usage. From Australia Government Smart City / Smart Grid Project (SCSG). Measured at 30-minute intervals across 223 households in southeastern Australia in 2010-2012. *Same data source as for reference series.* Temperature. Sydney airport database. Measured at 1-hour intervals at Sydney airport 2010-2012. Day of week. Generated by analysis platform. Hour of day (for day-ahead only). Generated by analysis platform.

6.4 Results

To facilitate comparison to results of our Ireland analysis, we again highlight results around MAPE-based techniques. Our discussion centers on results from week-ahead analysis, but we include data visualizations of results from day-ahead analysis, too.

6.4.1 Data Characterization

With Australia data captured at one-day resolution, as used for our week-ahead forecasting, evident is a more complex relationship between electricity usage and temperature than with Ireland data (Figure 6-3, Figure 6-4). Positive and negative correlations are present at different times of year corresponding to temperature peaks and troughs, respectively. The positive correlation at temperature peaks is somewhat less pronounced than the negative correlation at temperature troughs. Electricity usage increases for cooling when temperatures are high, increases for heating when temperatures are low, and decreases when temperatures are mild. (What constitutes temperatures low enough to trigger increases in electricity usage differs between Ireland and Australia.)

6.4.2 Sensitivity of Metric Scores to Decisions

Twenty-five out of 30 metrics considered yielded valid forecast accuracy evaluations for all forecasts.

Different decisions influence forecasting performance disproportionately (Figure 6-5). The span decision accounts for most relative importance for 4 out of 25 metrics, algorithm class for 6 metrics, extension rule for 10 metrics, and holdout for 5 metrics. As scored by MAPE, decisions for algorithm class and extension rule together account for about 85% relative importance; span accounts for little relative importance.

MAPE scores are distributed in a distinctive pattern similar to that as for Ireland, but with scores concentrated at 10% to 20% rather than 5% to 10% (Figure 6-7). Considered in the context of fixed algorithm class decisions, the best MAPE scores are distributed across

techniques using any of the six algorithms, all other decisions being equal (Figure 6-8). As for Ireland, MAPE score either increases or decreases as holdout increases depending on the technique and the holdout option being compared (Figure 6-10).

6.4.3 Metric Relationships

Scores arranged by technique are strongly correlated at > 0.9 in only 28 out of 300 pairs of metrics (Figure 6-11). As compared to Ireland, many more pairs are poorly correlated.

6.4.4 Sensitivity of Ranks to Decisions

Techniques ranked per MAPE show a distinctive decision effectiveness signature (Figure 6-13). The top-ranked technique per MAPE uses a 6-month span, linear regression algorithm, recursive extension rule, 1-day update cycle, 20% holdout, and 20% clip (Figure 6-14, Table 6-1). All of the 20 best performing techniques use the shortest span, the smallest holdout, and one of the two largest clips. Techniques ranked 2 through 5 differ only by update cycle. All algorithms, all update cycles, and both extension rules are represented among the 20 best performing techniques (Figure 6-15).

Looking more closely at effects of individual decisions on MAPE score, the best 13% of techniques that use a 6-month span beat all other techniques; 75% of techniques that use a 6-month span beat by all other techniques (Figure 6-19). Techniques that use a 12-month span and those that use an 18-month span perform about as well each other, when compared rank-to-rank. As noted earlier, all algorithms are represented among the best performing techniques, though as performance decreases, multilayer perceptron soon falls away and naïve later falls away (Figure 6-20). Except for the best performing techniques, those that use a recursive extension rule mostly dominate techniques that use a direct extension rule (Figure 6-21). Techniques that use shorter update cycles mostly dominate techniques that use longer update cycles, but not by much (Figure 6-22). Techniques that use the largest holdout mostly dominate the other techniques (Figure 6-23). Techniques that use a larger clip mostly dominate techniques that use a smaller clip, but not by much (Figure 6-24).

6.4.5 Sensitivity of Ranks to Decisions in Context

The stories sound different when ranks are considered in the context of locking-in certain decision options.

In the context of a decision to use a short span, choice of algorithm does not matter much among the best performing techniques (Figure 6-25). However, through a lens looking only at spans longer than 6 months, the algorithm decision distinguishes techniques considerably. At a 12-month span, no techniques that use multilayer perceptron or support vector regression are represented in the 13% best performing techniques (Figure 6-26). At an 18-month span, no techniques that use k-nearest neighbor, decision tree, or multilayer perceptron are represented among the 10% best performing techniques (Figure 6-27).

6.4.6 Sampling

Like in Ireland, MAPE scores vary widely across techniques applied to single household samples, but as sample size increases to only 5%-10% of population size, they converge quickly to tight distributions approximating the population distribution (Figure 6-41).

Also as seen with Ireland, techniques that perform well for one household do not generally perform very well for other households (Figure 6-43). As sample size increases, technique ranks within a sample, as a group, move monotonically to align with technique ranks within the population, though not as quickly as MAPE score moves (Figure 6-43, Figure 6-44).

6.4.7 Clustering

Like in Ireland, the number of clusters used to produce population forecasts has little effect on MAPE score distributions across techniques (Figure 6-47). As the number of clusters decreases, technique ranks within a cluster grouping, as a group, move monotonically to align with techniques ranks within the population, but individual technique ranks do so non-monotonically (Figure 6-49).

6.4.8 Temporal Magnification

Like in Ireland, MAPE score distributions across techniques at finer resolutions – 6-hour and 12-hour resolutions as compared to a 1-day resolution baseline – tend to worsen (Figure 6-51).

6.5 Insights

We glean the following insights from our results, with the requisite caveat that they are based on two specific smart electric grid datasets and a practically scoped set of experiments.

6.5.1 Training and Testing Decisions

> Training and testing decisions can dominate other forecasting process decisions.

The common thread running through Australia's best performing techniques is that they all use a small amount of testing data, the result of a short span, a small holdout, and a large clip. Anthropomorphizing freely here, Australia is leveraging training and testing decisions and a small portion of its data that happens to be easily forecastable by techniques using most any algorithm, extension rule, or update cycle. Comparison of techniques across Ireland and Australia is still legitimate because Ireland is afforded the same opportunity. It just turns out for Ireland that no small portion of easily forecastable data could be isolated by training and testing decisions.

6.5.2 Location-Specific Techniques

> Different locations lend themselves to different forecasting techniques.

Ireland's 20 best performing techniques have little in common with Australia's 20 best performing techniques (Table 6-3, Table 6-4). None of Ireland's are among Australia's, and none of Australia's are among Ireland's.

Ireland's 20 best performing techniques produce forecasts that evaluate to a mean MAPE of 2.69%. These same techniques applied in Australia produce forecasts that evaluate to a mean MAPE of 10.90%, with mean rank 718th, ranging from 21st to 1,726th out of 2,880. Ireland's best technique applied in Australia would produce a forecast 9.46 MAPE points, or 6.41 times, worse than one produced by Australia's best technique. Across all of Ireland's best techniques applied in Australia, they would on average produce a forecast 7.38 MAPE points, or 2.30 times, worse than those produced by Australia's best techniques.

Looking at it the other way around, Australia's 20 best performing techniques applied in Ireland produce among the very worst forecasts, evaluating to a mean MAPE of 29.41%, with mean rank 2,723rd, ranging from 1,389th to 2,877th. Australia's best technique applied in Ireland would produce a forecast 8.93 MAPE points, or 3.83 times, worse than one produced by Ireland's bets technique. Across all of Australia's best techniques applied in Ireland, they would on average produce a forecast 26.71 MAPE points, or 9.91 times, worse than those produced by Ireland's best techniques.

In light of what we learned about Australia's fondness for small amounts of testing data, we additionally compared Ireland and Australia results excluding forecasts that use a 6-month span, to see if the dissimilarity of best performing techniques persists (Table 6-5, Table 6-6). Through this lens, two of Ireland's 20 best performing techniques are represented among Australia's: Ireland's 18th is Australia's 8th, Ireland's 19th is Australia's 1st. Ireland's 20 best applied in Australia produce forecasts that evaluate to a mean MAPE of 11.33%, with mean rank 511th, ranging from 1st to 1,270th. Australia's 20 best applied in Ireland produce forecasts that evaluate to a mean MAPE of 3.33%, with mean rank 111th, ranging from 18th to 349th. Australia's techniques work better in Ireland than Ireland's do in Australia. There is less, but still significant, dissimilarity when forecasts using a 6-month span are excluded.

6.5.3 "One-Size-Fits-All" Techniques

> Forecasting performance at different locations is not much degraded by "one-size-fits-all" techniques.

To find the set of techniques that optimally compromise the MAPE impact across multiple locations, we rank per the order of the sums of individual location ranks (Table 6-7). We call the 20 best techniques so ranked the "one-size-fits-all" techniques.

The "one-size-fits-all" techniques produce forecasts that evaluate to better MAPE scores than do Ireland's best applied to Australia or Australia's best applied to Ireland, with an Ireland mean MAPE of 3.09% and an Australia mean MAPE of 7.54%. The best "one-size-fits-all" technique applied in Ireland would produce a forecast 0.18 MAPE points, or 0.08 times, worse than one produced by Ireland's best. The best "one-size-fits-all" technique applied in Australia would produce a forecast 3.12 MAPE points, or 2.11 times worse than one produced by Australia's best. Across all the "one-size-fits-all" techniques applied in Ireland, they would on average produce a forecast 0.40 MAPE points, or 0.15 times, worse than those produced by Ireland's best techniques. Across all the "one-size-fits-all" techniques applied in Australia, they would on average produce a forecast 4.02 MAPE points, or 1.21 times, worse than those produced by Australia's best techniques. Mean impacts of 0.40 MAPE points (in Ireland) and 4.02 MAPE points (in Australia) may or may not be considered acceptable. By this standard, one set of "one-size-fits-all" techniques may or may not be a reasonable substitute for many sets of location-specific best techniques.

Again, we additionally compared Ireland and Australia results excluding forecasts that use a 6-month span, to see if the level of impact persists (Table 6-8). In this case, the mean impact on Ireland would be 0.34 MAPE points, or 0.13 times, worse. The impact on Australia would be 0.37 MAPE points, or 0.05 times, worse. Mean impacts of 0.34 MAPE points (in Ireland) and 0.37 MAPE points (in Australia) likely may be considered acceptable. By this standard, one set

of “one-size-fits-all” techniques may be a reasonable substitute for many sets of location-specific best techniques.

6.5.4 Sampling

> Forecasting performance at different locations is not much degraded by sampling.

For each location, we can formulate a relationship between MAPE impact and sample size, where impact is expressed as a third-degree polynomial of the logarithm of the sample size normalized with respect to the population size, adjusted proportionally by proximity to full population size to force intersection with the mean impact at full population size. From this relationship, we can estimate the impact for some given sample size (Figure 6-53), or conversely, estimate the minimum sample size that restricts impact to within some given tolerance (Figure 6-54). For example, for Ireland, techniques applied per a data strategy assuming a sample 11.2% or more of the population size are estimated on average to produce forecasts that evaluate to within 2.5 MAPE points of the mean for techniques applied without this data strategy. For Australia, a sample of 19.3% or more will do it.

The shapes of the impact/sample size curves match closely across locations. In both cases, we see that use of even a small sample size (say anything larger than 5%-20% of population size) does not much degrade forecasting performance.

6.5.5 Clustering

> Forecasting performance at different locations is not much improved by clustering.

For each location, we can formulate a relationship between MAPE impact and number of clusters, where impact is expressed as a third-degree polynomial of the logarithm of the number of clusters, adjusted proportionally by proximity to no clustering to force intersection with the mean impact at no clustering.

From this relationship, we can estimate the impact for some given number of clusters (Figure 6-57), or conversely, estimate the maximum number of clusters that restricts impact to within some given tolerance (Figure 6-58). For example, for Ireland, techniques applied per a data strategy assuming any number of clusters are estimated on average to produce forecasts that evaluate to within 0.5 MAPE points of the mean for techniques applied without this data strategy. For Australia, about 12 clusters or fewer will do it.

The shapes of the impact/number of clusters curves match closely across locations. In both cases, we see that even a wide variety in number of clusters does not much improve forecasting performance.

6.5.6 Temporal Magnification

> Forecasting performance at different locations is degraded by refining temporal magnification at coarse time step sizes.

For each location, we can formulate a relationship between MAPE impact and time step size, where impact is expressed as a second-degree polynomial of the time step size, over the range 6 hours to 24 hours, in the context of week-ahead forecasts.

From this relationship, we can estimate the impact for some given time step size (Figure 6-60), or conversely, estimate the minimum time step size that restricts impact to within some given tolerance (Figure 6-61). For example, for Ireland, techniques applied per a data strategy assuming time step size of 19 hours or more are estimated on average to produce forecasts that evaluate to within 5 MAPE points of the mean for techniques applied without this data strategy. For Australia, time step sizes of 20 hours or more will do it.

The shapes of the impact/time step size curves match closely across locations. In both cases, we see that even a small decrease in time step size degrades forecasting performance, over the range 6 hours to 24 hours.

6.6 Tables and Data Visualizations

Table 6-1: Best performing techniques, ranked by MAPE score, Australia week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	MAPE.r
168	182	24	linreg	recurse	24	0.20	0.20	1.48	1
168	182	24	knn	recurse	48	0.20	0.20	2.42	2
168	182	24	knn	recurse	72	0.20	0.20	2.42	3
168	182	24	knn	recurse	120	0.20	0.20	2.69	4
168	182	24	knn	recurse	168	0.20	0.20	2.69	5
168	182	24	svm	direct	48	0.20	0.10	2.92	6
168	182	24	svm	direct	24	0.20	0.10	3.46	7
168	182	24	svm	direct	120	0.20	0.10	3.51	8
168	182	24	svm	direct	168	0.20	0.10	3.51	9
168	182	24	mlp	direct	120	0.20	0.10	3.59	10
168	182	24	mlp	direct	168	0.20	0.10	3.59	11
168	182	24	naive	recurse	72	0.20	0.10	3.69	12
168	182	24	svm	direct	24	0.20	0.20	4.07	13
168	182	24	svm	recurse	120	0.20	0.10	4.11	14
168	182	24	svm	recurse	168	0.20	0.10	4.21	15
168	182	24	knn	direct	24	0.20	0.10	4.21	16
168	182	24	mlp	recurse	120	0.20	0.20	4.37	17
168	182	24	knn	recurse	24	0.20	0.20	4.40	18
168	182	24	tree	recurse	120	0.20	0.10	4.44	19
168	182	24	tree	direct	24	0.20	0.10	4.47	20
							mean	3.51	10.50

Table 6-2: Best performing techniques, ranked by MAPE score, Australia day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	MAPE	MAPE.r
24	183	1	naive	recurse	1	0.20	0.20	8.98	1
24	183	1	naive	recurse	1	0.20	0.10	9.03	2
24	183	1	naive	recurse	1	0.33	0.10	9.05	3
24	183	1	linreg	recurse	1	0.20	0.20	9.06	4
24	183	1	naive	recurse	1	0.33	0.20	9.08	5
24	183	1	linreg	recurse	1	0.20	0.10	9.11	6
24	183	1	naive	recurse	1	0.33	0	9.17	7
24	183	1	naive	recurse	1	0.20	0	9.23	8
24	183	1	linreg	recurse	1	0.20	0	9.29	9
24	363	1	naive	recurse	1	0.20	0.10	9.34	10
24	183	1	linreg	recurse	1	0.33	0.10	9.37	11
24	183	1	linreg	recurse	1	0.33	0.20	9.40	12
24	183	1	mlp	recurse	1	0.20	0.10	9.40	13
24	363	1	naive	recurse	1	0.20	0	9.41	14
24	363	1	naive	recurse	1	0.20	0.20	9.43	15
24	183	1	linreg	recurse	1	0.33	0	9.47	16
24	183	1	mlp	recurse	1	0.20	0.20	9.57	17
24	183	1	naive	recurse	1	0.50	0.10	9.63	18
24	183	1	naive	recurse	1	0.50	0	9.65	19
24	183	1	naive	recurse	1	0.50	0.20	9.66	20
							mean	9.32	10.50

Table 6-3: Impact on Australia when Ireland best techniques are used, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	aus_impact	aus_impact_per
168	363	24	naive	direct	24	0.50	0.20	10.94	470	2.33	1	9.46	6.41
168	534	24	knn	recurse	24	0.30	0.20	11.86	768	2.34	2	9.43	3.89
168	534	24	naive	direct	24	0.40	0.20	10.33	337	2.47	3	7.91	3.26
168	534	24	linreg	direct	24	0.30	0.20	13.19	1314	2.47	4	10.50	3.90
168	182	24	naive	recurse	48	0.20	0.20	4.60	21	2.51	5	1.91	0.71
168	182	24	naive	recurse	72	0.20	0.20	4.60	22	2.51	6	1.68	0.57
168	534	24	linreg	direct	24	0.40	0.20	11.95	797	2.55	7	8.49	2.46
168	534	24	naive	direct	24	0.30	0.20	13.56	1477	2.57	8	10.05	2.87
168	534	24	knn	recurse	24	0.30	0.10	12.24	903	2.72	9	8.73	2.49
168	534	24	tree	recurse	24	0.30	0.20	11.05	494	2.76	10	7.46	2.08
168	534	24	knn	recurse	48	0.30	0.20	11.75	735	2.78	11	8.16	2.27
168	534	24	naive	direct	24	0.50	0.20	10.37	346	2.79	12	6.69	1.81
168	534	24	tree	recurse	24	0.40	0.20	9.78	271	2.82	13	5.71	1.40
168	534	24	tree	recurse	24	0.40	0.10	10.35	341	2.83	14	6.25	1.52
168	534	24	knn	recurse	72	0.30	0.20	12.83	1156	2.85	15	8.63	2.05
168	534	24	knn	recurse	24	0.40	0.20	9.66	254	2.87	16	5.45	1.29
168	534	24	knn	direct	24	0.30	0.20	14.05	1663	2.91	17	9.68	2.21
168	534	24	naive	direct	24	0.40	0.10	12.86	1165	2.92	18	8.47	1.93
168	534	24	knn	direct	48	0.30	0.20	14.18	1726	2.94	19	9.74	2.19
168	534	24	linreg	recurse	24	0.40	0.20	7.77	101	2.94	20	3.29	0.74
							mean	10.90	718.05	2.69	10.50	7.38	2.30

Table 6-4: Impact on Ireland when Australia best techniques are used, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	cer_impact	cer_impact_per
168	182	24	linreg	recurse	24	0.20	0.20	1.48	1	11.26	1880	8.93	3.83
168	182	24	knn	recurse	48	0.20	0.20	2.42	2	27.95	2800	25.61	10.95
168	182	24	knn	recurse	72	0.20	0.20	2.42	3	27.95	2801	25.48	10.32
168	182	24	knn	recurse	120	0.20	0.20	2.69	4	27.95	2802	25.47	10.29
168	182	24	knn	recurse	168	0.20	0.20	2.69	5	28.20	2803	25.69	10.22
168	182	24	svm	direct	48	0.20	0.10	2.92	6	33.42	2868	30.90	12.29
168	182	24	svm	direct	24	0.20	0.10	3.46	7	33.20	2863	30.65	12.03
168	182	24	svm	direct	120	0.20	0.10	3.51	8	34.52	2876	31.94	12.41
168	182	24	svm	direct	168	0.20	0.10	3.51	9	34.52	2877	31.80	11.69
168	182	24	mlp	direct	120	0.20	0.10	3.59	10	33.00	2856	30.23	10.95
168	182	24	mlp	direct	168	0.20	0.10	3.59	11	33.00	2857	30.22	10.89
168	182	24	naive	recurse	72	0.20	0.10	3.69	12	7.63	1389	4.84	1.74
168	182	24	svm	direct	24	0.20	0.20	4.07	13	30.17	2824	27.35	9.69
168	182	24	svm	recurse	120	0.20	0.10	4.11	14	34.32	2874	31.49	11.14
168	182	24	svm	recurse	168	0.20	0.10	4.21	15	34.03	2873	31.18	10.95
168	182	24	knn	direct	24	0.20	0.10	4.21	16	31.13	2841	28.25	9.84
168	182	24	mlp	recurse	120	0.20	0.20	4.37	17	30.57	2837	27.66	9.50
168	182	24	knn	recurse	24	0.20	0.20	4.40	18	30.00	2823	27.08	9.28
168	182	24	tree	recurse	120	0.20	0.10	4.44	19	33.36	2866	30.42	10.36
168	182	24	tree	direct	24	0.20	0.10	4.47	20	31.97	2846	29.03	9.87
							mean	3.51	10.50	29.41	2722.80	26.71	9.91

Table 6-5:

Impact on Australia when Ireland best techniques are used,
no 6-month span, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	aus_impact	aus_impact_per
168	363	24	naive	direct	24	0.50	0.20	10.94	263	2.33	1	4.53	0.71
168	534	24	knn	recurse	24	0.30	0.20	11.86	530	2.34	2	4.94	0.71
168	534	24	naive	direct	24	0.40	0.20	10.33	160	2.47	3	3.14	0.44
168	534	24	linreg	direct	24	0.30	0.20	13.19	984	2.47	4	5.87	0.80
168	534	24	linreg	direct	24	0.40	0.20	11.95	556	2.55	5	4.59	0.62
168	534	24	naive	direct	24	0.30	0.20	13.56	1094	2.57	6	6.07	0.81
168	534	24	knn	recurse	24	0.30	0.10	12.24	654	2.72	7	4.69	0.62
168	534	24	tree	recurse	24	0.30	0.20	11.05	284	2.76	8	3.29	0.42
168	534	24	knn	recurse	48	0.30	0.20	11.75	500	2.78	9	3.97	0.51
168	534	24	naive	direct	24	0.50	0.20	10.37	166	2.79	10	2.52	0.32
168	534	24	tree	recurse	24	0.40	0.20	9.78	112	2.82	11	1.92	0.24
168	534	24	tree	recurse	24	0.40	0.10	10.35	162	2.83	12	2.33	0.29
168	534	24	knn	recurse	72	0.30	0.20	12.83	866	2.85	13	4.59	0.56
168	534	24	knn	recurse	24	0.40	0.20	9.66	100	2.87	14	1.39	0.17
168	534	24	knn	direct	24	0.30	0.20	14.05	1224	2.91	15	5.75	0.69
168	534	24	naive	direct	24	0.40	0.10	12.86	873	2.92	16	4.55	0.55
168	534	24	knn	direct	48	0.30	0.20	14.18	1270	2.94	17	5.85	0.70
168	534	24	linreg	recurse	24	0.40	0.20	7.77	8	2.94	18	-0.73	-0.09
168	534	24	naive	recurse	24	0.50	0.20	6.41	1	2.94	19	-2.11	-0.25
168	363	24	naive	direct	48	0.50	0.20	11.48	414	2.96	20	2.96	0.35
							mean	11.33	511.05	2.74	10.50	3.51	0.46

Table 6-6:

Impact on Ireland when Australia best techniques are used,
no 6-month span, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	cer_impact	cer_impact_per
168	534	24	naive	recurse	24	0.50	0.20	6.41	1	2.94	19	0.61	0.26
168	534	24	naive	recurse	24	0.40	0.20	6.91	2	2.97	24	0.63	0.27
168	534	24	naive	recurse	24	0.50	0.10	7.19	3	2.99	30	0.52	0.21
168	534	24	naive	recurse	24	0.50	0.01	7.32	4	3.19	54	0.71	0.29
168	534	24	naive	recurse	24	0.50	0	7.36	5	3.20	58	0.65	0.25
168	534	24	naive	recurse	72	0.40	0.20	7.49	6	3.97	302	1.39	0.54
168	534	24	naive	recurse	48	0.50	0.20	7.55	7	3.49	136	0.78	0.29
168	534	24	linreg	recurse	24	0.40	0.20	7.77	8	2.94	18	0.18	0.07
168	534	24	naive	recurse	24	0.40	0.10	7.78	9	2.97	25	0.20	0.07
168	534	24	naive	recurse	24	0.40	0	7.86	10	3.27	76	0.48	0.17
168	534	24	naive	recurse	24	0.40	0.01	7.86	11	3.27	75	0.45	0.16
168	363	24	linreg	recurse	24	0.50	0.20	8.02	12	3.36	109	0.54	0.19
168	534	24	naive	recurse	48	0.40	0.20	8.24	13	3.71	203	0.87	0.30
168	363	24	linreg	recurse	24	0.50	0.10	8.27	14	3.68	194	0.81	0.28
168	363	24	linreg	recurse	24	0.50	0	8.31	15	4.08	339	1.17	0.40
168	363	24	naive	recurse	24	0.50	0.20	8.32	16	3.01	34	0.09	0.03
168	363	24	linreg	recurse	24	0.50	0.01	8.33	17	4.10	349	1.16	0.40
168	534	24	linreg	recurse	24	0.40	0	8.50	18	3.21	60	0.26	0.09
168	363	24	naive	recurse	24	0.50	0	8.51	19	3.36	107	0.42	0.14
168	534	24	linreg	recurse	24	0.40	0.10	8.52	20	2.96	21	0.00	0.00
							mean	7.83	10.50	3.33	111.65	0.60	0.22

Table 6-7:
Impact on Australia and Ireland when "one-size-fits-all" best techniques are used, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r
168	182	24	naive	recurse	48	0.20	0.20	4.60	21	2.51	5
168	182	24	naive	recurse	72	0.20	0.20	4.60	22	2.51	6
168	534	24	naive	recurse	24	0.50	0.20	6.41	61	2.94	21
168	534	24	naive	recurse	24	0.40	0.20	6.91	77	2.97	26
168	534	24	naive	recurse	24	0.50	0.10	7.19	81	2.99	32
168	534	24	linreg	recurse	24	0.40	0.20	7.77	101	2.94	20
168	534	24	naive	recurse	24	0.40	0.10	7.78	102	2.97	27
168	534	24	naive	recurse	24	0.50	0.01	7.32	83	3.19	56
168	534	24	naive	recurse	24	0.50	0	7.36	89	3.20	60
168	534	24	linreg	recurse	24	0.40	0.10	8.52	136	2.96	23
168	363	24	naive	recurse	24	0.50	0.20	8.32	129	3.01	36
168	363	24	naive	recurse	24	0.50	0.10	8.62	142	3.05	40
168	534	24	naive	recurse	24	0.40	0	7.86	105	3.27	79
168	534	24	naive	recurse	24	0.40	0.01	7.86	106	3.27	78
168	534	24	linreg	recurse	24	0.40	0	8.50	134	3.21	62
168	534	24	linreg	recurse	24	0.40	0.01	8.53	138	3.23	71
168	363	24	linreg	recurse	24	0.50	0.20	8.02	115	3.36	113
168	534	24	naive	recurse	48	0.50	0.20	7.55	94	3.49	143
168	363	24	naive	recurse	24	0.50	0	8.51	135	3.36	111
168	363	24	naive	recurse	24	0.50	0.01	8.52	137	3.36	112
							mean	7.54	100.40	3.09	56.05

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	ceraus_MAPE.r	aus_impact	aus_impact_per	cer_impact	cer_impact_per	ceraus_impact
168	182	24	naive	recurse	48	0.20	0.20	1	3.12	2.11	0.18	0.08	3.31
168	182	24	naive	recurse	72	0.20	0.20	2	2.17	0.90	0.18	0.08	2.35
168	534	24	naive	recurse	24	0.50	0.20	3	3.98	1.64	0.47	0.19	4.45
168	534	24	naive	recurse	24	0.40	0.20	4	4.22	1.57	0.50	0.20	4.72
168	534	24	naive	recurse	24	0.50	0.10	5	4.50	1.67	0.48	0.19	4.97
168	534	24	linreg	recurse	24	0.40	0.20	6	4.85	1.66	0.43	0.17	5.27
168	534	24	naive	recurse	24	0.40	0.10	7	4.32	1.25	0.43	0.17	4.75
168	534	24	naive	recurse	24	0.50	0.01	8	3.82	1.09	0.61	0.24	4.43
168	534	24	naive	recurse	24	0.50	0	9	3.86	1.10	0.48	0.18	4.33
168	534	24	linreg	recurse	24	0.40	0.10	10	4.92	1.37	0.19	0.07	5.12
168	363	24	naive	recurse	24	0.50	0.20	11	4.72	1.31	0.24	0.08	4.96
168	363	24	naive	recurse	24	0.50	0.10	12	4.94	1.34	0.26	0.09	5.20
168	534	24	naive	recurse	24	0.40	0	13	3.78	0.93	0.45	0.16	4.23
168	534	24	naive	recurse	24	0.40	0.01	14	3.75	0.91	0.44	0.16	4.19
168	534	24	linreg	recurse	24	0.40	0	15	4.29	1.02	0.36	0.13	4.65
168	534	24	linreg	recurse	24	0.40	0.01	16	4.32	1.02	0.36	0.13	4.68
168	363	24	linreg	recurse	24	0.50	0.20	17	3.65	0.83	0.45	0.16	4.10
168	534	24	naive	recurse	48	0.50	0.20	18	3.16	0.72	0.58	0.20	3.73
168	363	24	naive	recurse	24	0.50	0	19	4.07	0.92	0.42	0.14	4.49
168	363	24	naive	recurse	24	0.50	0.01	20	4.05	0.91	0.42	0.14	4.47
							mean	10.50	4.02	1.21	0.40	0.15	4.42

Table 6-8:

Impact on Australia and Ireland when "one-size-fits-all" best techniques are used, no 6-month span, week-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r
168	534	24	naive	recurse	24	0.50	0.20	6.41	1	2.94	19
168	534	24	linreg	recurse	24	0.40	0.20	7.77	8	2.94	18
168	534	24	naive	recurse	24	0.40	0.20	6.91	2	2.97	24
168	534	24	naive	recurse	24	0.50	0.10	7.19	3	2.99	30
168	534	24	naive	recurse	24	0.40	0.10	7.78	9	2.97	25
168	534	24	linreg	recurse	24	0.40	0.10	8.52	20	2.96	21
168	363	24	naive	recurse	24	0.50	0.20	8.32	16	3.01	34
168	534	24	naive	recurse	24	0.50	0.01	7.32	4	3.19	54
168	363	24	naive	recurse	24	0.50	0.10	8.62	25	3.05	38
168	534	24	naive	recurse	24	0.50	0	7.36	5	3.20	58
168	534	24	linreg	recurse	24	0.40	0	8.50	18	3.21	60
168	534	24	naive	recurse	24	0.40	0	7.86	10	3.27	76
168	534	24	naive	recurse	24	0.40	0.01	7.86	11	3.27	75
168	534	24	linreg	recurse	24	0.40	0.01	8.53	22	3.23	68
168	534	24	linreg	recurse	24	0.20	0.20	9.43	76	2.99	29
168	534	24	knn	recurse	24	0.40	0.20	9.66	100	2.87	14
168	363	24	linreg	recurse	24	0.50	0.20	8.02	12	3.36	109
168	534	24	tree	recurse	24	0.40	0.20	9.78	112	2.82	11
168	363	24	naive	recurse	24	0.50	0	8.51	19	3.36	107
168	363	24	naive	recurse	24	0.30	0.20	9.63	98	2.99	31
							mean	8.20	28.55	3.08	45.05

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	ceraus_MAPE.r	aus_impact	aus_impact_per	cer_impact	cer_impact_per	ceraus_impact
168	534	24	naive	recurse	24	0.50	0.20	1	0	0	0.61	0.26	0.61
168	534	24	linreg	recurse	24	0.40	0.20	2	0.85	0.12	0.60	0.26	1.45
168	534	24	naive	recurse	24	0.40	0.20	3	-0.28	-0.04	0.50	0.20	0.23
168	534	24	naive	recurse	24	0.50	0.10	4	-0.13	-0.02	0.51	0.21	0.38
168	534	24	naive	recurse	24	0.40	0.10	5	0.42	0.06	0.43	0.17	0.84
168	534	24	linreg	recurse	24	0.40	0.10	6	1.03	0.14	0.38	0.15	1.41
168	363	24	naive	recurse	24	0.50	0.20	7	0.77	0.10	0.29	0.11	1.06
168	534	24	naive	recurse	24	0.50	0.01	8	-0.44	-0.06	0.43	0.15	-0.02
168	363	24	naive	recurse	24	0.50	0.10	9	0.84	0.11	0.28	0.10	1.12
168	534	24	naive	recurse	24	0.50	0	10	-0.49	-0.06	0.41	0.15	-0.08
168	534	24	linreg	recurse	24	0.40	0	11	0.64	0.08	0.38	0.14	1.02
168	534	24	naive	recurse	24	0.40	0	12	-0.17	-0.02	0.44	0.16	0.28
168	534	24	naive	recurse	24	0.40	0.01	13	-0.38	-0.05	0.42	0.15	0.04
168	534	24	linreg	recurse	24	0.40	0.01	14	0.25	0.03	0.36	0.13	0.61
168	534	24	linreg	recurse	24	0.20	0.20	15	1.12	0.13	0.08	0.03	1.20
168	534	24	knn	recurse	24	0.40	0.20	16	1.35	0.16	-0.04	-0.02	1.30
168	363	24	linreg	recurse	24	0.50	0.20	17	-0.31	-0.04	0.43	0.15	0.12
168	534	24	tree	recurse	24	0.40	0.20	18	1.28	0.15	-0.12	-0.04	1.17
168	363	24	naive	recurse	24	0.50	0	19	0	0	0.42	0.14	0.42
168	363	24	naive	recurse	24	0.30	0.20	20	1.11	0.13	0.04	0.01	1.15
							mean	10.50	0.37	0.05	0.34	0.13	0.72

Table 6-9: Impact on Australia when Ireland best techniques are used, day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	aus_impact	aus_impact_per
24	535	1	linreg	direct	1	0.33	0.20	10.43	40	4.90	1	1.44	0.16
24	535	1	linreg	direct	1	0.33	0.10	10.97	52	5.10	2	1.95	0.22
24	535	1	linreg	direct	1	0.50	0.20	9.72	21	5.12	3	0.67	0.07
24	363	1	linreg	direct	1	0.50	0.20	10.26	36	5.13	4	1.20	0.13
24	535	1	linreg	direct	1	0.50	0.10	10.41	39	5.17	5	1.33	0.15
24	363	1	linreg	direct	1	0.50	0.10	10.66	43	5.32	6	1.55	0.17
24	535	1	linreg	direct	1	0.20	0.20	12.56	106	5.46	7	3.38	0.37
24	363	1	linreg	direct	1	0.33	0.20	12.11	90	5.66	8	2.87	0.31
24	535	1	linreg	direct	1	0.20	0.10	12.31	96	5.71	9	3.03	0.33
24	535	1	linreg	direct	1	0.50	0	10.70	46	5.71	10	1.36	0.15
24	363	1	linreg	direct	1	0.33	0.10	12.11	91	5.79	11	2.74	0.29
24	183	1	linreg	direct	1	0.50	0.20	12.28	94	5.80	12	2.89	0.31
24	183	1	linreg	direct	1	0.33	0.20	14.74	184	5.92	13	5.34	0.57
24	535	1	linreg	direct	1	0.33	0	11.14	60	5.93	14	1.74	0.18
24	363	1	linreg	direct	1	0.50	0	10.72	47	6.09	15	1.29	0.14
24	183	1	linreg	direct	1	0.50	0.10	12.35	98	6.22	16	2.89	0.31
24	363	1	linreg	direct	1	0.20	0.20	13.37	139	6.43	17	3.80	0.40
24	363	1	linreg	direct	1	0.33	0	12.02	85	6.85	18	2.39	0.25
24	535	1	linreg	direct	1	0.20	0	12.48	103	6.88	19	2.83	0.29
24	183	1	linreg	direct	1	0.33	0.10	14.79	189	7.02	20	5.13	0.53
							mean	11.81	82.95	5.81	10.50	2.49	0.27

Table 6-10: Impact on Ireland when Australia best techniques are used, day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r	cer_impact	cer_impact_per
24	183	1	naive	recurse	1	0.20	0.20	8.98	1	13.73	114	8.83	1.80
24	183	1	naive	recurse	1	0.20	0.10	9.03	2	13.43	112	8.32	1.63
24	183	1	naive	recurse	1	0.33	0.10	9.05	3	14.47	130	9.36	1.83
24	183	1	linreg	recurse	1	0.20	0.20	9.06	4	13.53	113	8.40	1.64
24	183	1	naive	recurse	1	0.33	0.20	9.08	5	14.72	137	9.55	1.85
24	183	1	linreg	recurse	1	0.20	0.10	9.11	6	13.27	109	7.95	1.49
24	183	1	naive	recurse	1	0.33	0	9.17	7	14.25	122	8.78	1.61
24	183	1	naive	recurse	1	0.20	0	9.23	8	13.37	111	7.71	1.36
24	183	1	linreg	recurse	1	0.20	0	9.29	9	13.32	110	7.61	1.33
24	363	1	naive	recurse	1	0.20	0.10	9.34	10	14.63	134	8.92	1.56
24	183	1	linreg	recurse	1	0.33	0.10	9.37	11	14.22	121	8.43	1.46
24	183	1	linreg	recurse	1	0.33	0.20	9.40	12	14.36	126	8.57	1.48
24	183	1	mlp	recurse	1	0.20	0.10	9.40	13	18.97	229	13.04	2.20
24	363	1	naive	recurse	1	0.20	0	9.41	14	14.37	128	8.44	1.42
24	363	1	naive	recurse	1	0.20	0.20	9.43	15	14.86	142	8.77	1.44
24	183	1	linreg	recurse	1	0.33	0	9.47	16	14.06	118	7.84	1.26
24	183	1	mlp	recurse	1	0.20	0.20	9.57	17	18.47	219	12.04	1.87
24	183	1	naive	recurse	1	0.50	0.10	9.63	18	14.84	141	7.99	1.17
24	183	1	naive	recurse	1	0.50	0	9.65	19	14.55	132	7.68	1.12
24	183	1	naive	recurse	1	0.50	0.20	9.66	20	14.89	144	7.88	1.12
							mean	9.32	10.50	14.62	134.60	8.81	1.53

Table 6-11:
Impact on Australia and Ireland when "one-size-fits-all" best techniques are used, day-ahead forecasts.

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	aus_MAPE	aus_MAPE.r	cer_MAPE	cer_MAPE.r
24	535	1	linreg	direct	1	0.50	0.20	9.72	21	5.12	3
24	363	1	linreg	direct	1	0.50	0.20	10.26	36	5.13	4
24	535	1	linreg	direct	1	0.33	0.20	10.43	40	4.90	1
24	535	1	linreg	direct	1	0.50	0.10	10.41	39	5.17	5
24	363	1	linreg	direct	1	0.50	0.10	10.66	43	5.32	6
24	535	1	linreg	direct	1	0.33	0.10	10.97	52	5.10	2
24	535	1	linreg	direct	1	0.50	0	10.70	46	5.71	10
24	363	1	linreg	direct	1	0.50	0	10.72	47	6.09	15
24	535	1	naive	direct	1	0.50	0.20	10.44	41	7.37	25
24	535	1	linreg	direct	1	0.33	0	11.14	60	5.93	14
24	535	1	naive	direct	1	0.50	0.10	11.01	56	7.38	27
24	535	1	svm	direct	1	0.50	0.20	10.98	54	7.47	32
24	535	1	naive	direct	1	0.50	0	10.99	55	7.54	34
24	363	1	linreg	direct	1	0.33	0.20	12.11	90	5.66	8
24	363	1	linreg	direct	1	0.33	0.10	12.11	91	5.79	11
24	363	1	linreg	direct	1	0.33	0	12.02	85	6.85	18
24	363	1	naive	direct	1	0.50	0.20	11.54	73	7.46	30
24	535	1	naive	direct	1	0.33	0.10	11.68	77	7.38	26
24	535	1	naive	direct	1	0.33	0.20	11.56	74	7.40	29
24	363	1	naive	direct	1	0.50	0.10	11.66	76	7.39	28
							mean	11.06	57.80	6.31	16.40

lookahead.h	span.d	resolution.h	algorithm	strategy	update.h	holdout	clip	ceraus_MAPE.r	aus_impact	aus_impact_per	cer_impact	cer_impact_per	ceraus_impact
24	535	1	linreg	direct	1	0.50	0.20	1	0.74	0.08	0.21	0.04	0.95
24	363	1	linreg	direct	1	0.50	0.20	2	1.23	0.14	0.03	0.01	1.26
24	535	1	linreg	direct	1	0.33	0.20	3	1.38	0.15	-0.21	-0.04	1.16
24	535	1	linreg	direct	1	0.50	0.10	4	1.35	0.15	0.03	0.01	1.38
24	363	1	linreg	direct	1	0.50	0.10	5	1.57	0.17	0.16	0.03	1.73
24	535	1	linreg	direct	1	0.33	0.10	6	1.86	0.20	-0.22	-0.04	1.64
24	535	1	linreg	direct	1	0.50	0	7	1.53	0.17	0.25	0.05	1.78
24	363	1	linreg	direct	1	0.50	0	8	1.49	0.16	0.43	0.08	1.92
24	535	1	naive	direct	1	0.50	0.20	9	1.16	0.12	1.65	0.29	2.81
24	535	1	linreg	direct	1	0.33	0	10	1.80	0.19	0.21	0.04	2.02
24	535	1	naive	direct	1	0.50	0.10	11	1.64	0.17	1.59	0.27	3.23
24	535	1	svm	direct	1	0.50	0.20	12	1.58	0.17	1.67	0.29	3.26
24	535	1	naive	direct	1	0.50	0	13	1.58	0.17	1.61	0.27	3.20
24	363	1	linreg	direct	1	0.33	0.20	14	2.70	0.29	-0.27	-0.05	2.43
24	363	1	linreg	direct	1	0.33	0.10	15	2.68	0.28	-0.30	-0.05	2.38
24	363	1	linreg	direct	1	0.33	0	16	2.55	0.27	0.63	0.10	3.18
24	363	1	naive	direct	1	0.50	0.20	17	1.97	0.21	1.03	0.16	3.00
24	535	1	naive	direct	1	0.33	0.10	18	2.05	0.21	0.53	0.08	2.59
24	535	1	naive	direct	1	0.33	0.20	19	1.91	0.20	0.52	0.08	2.43
24	363	1	naive	direct	1	0.50	0.10	20	2.00	0.21	0.37	0.05	2.37
							mean	10.50	1.74	0.19	0.50	0.08	2.24

https://relentlesslife.org/sydney-australia-at-night-2/

Figure 6-1: Sydney, Australia.

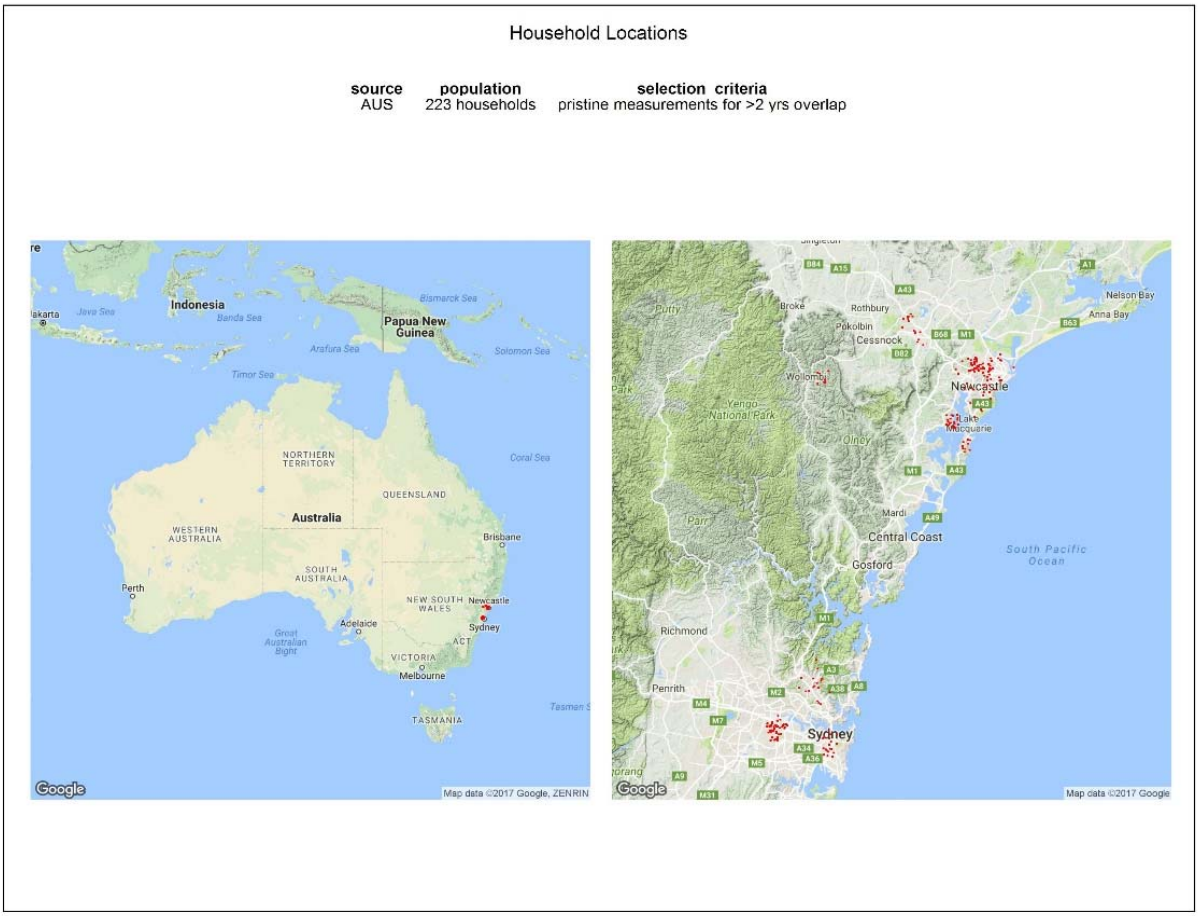


Figure 6-2: Australia household sites. 223 households. *Red* is household location.

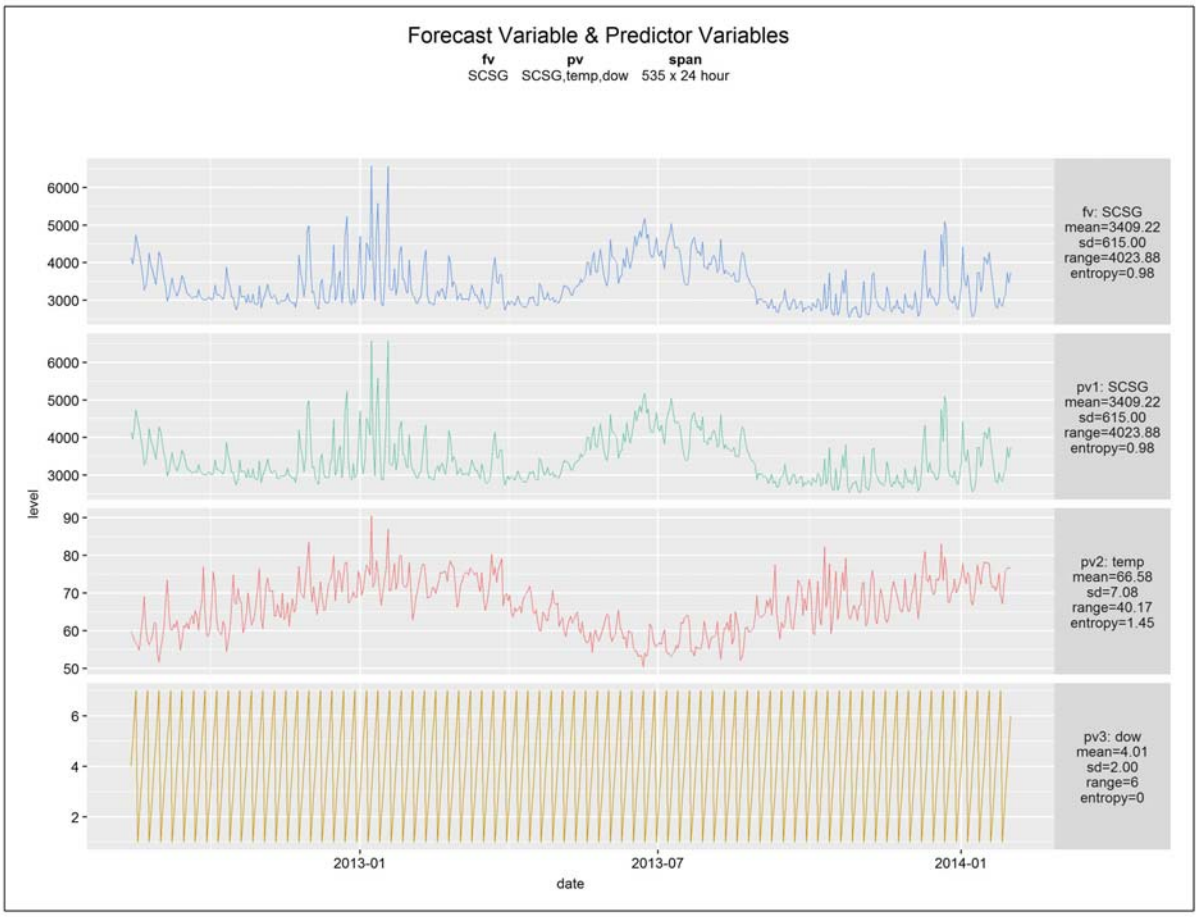


Figure 6-3: Reference and predictor series, Australia, for week-ahead forecasts. Reference series is aggregated 223-household electricity usage. Predictor series are aggregate electricity usage, temperature, and day of week (integer coded).

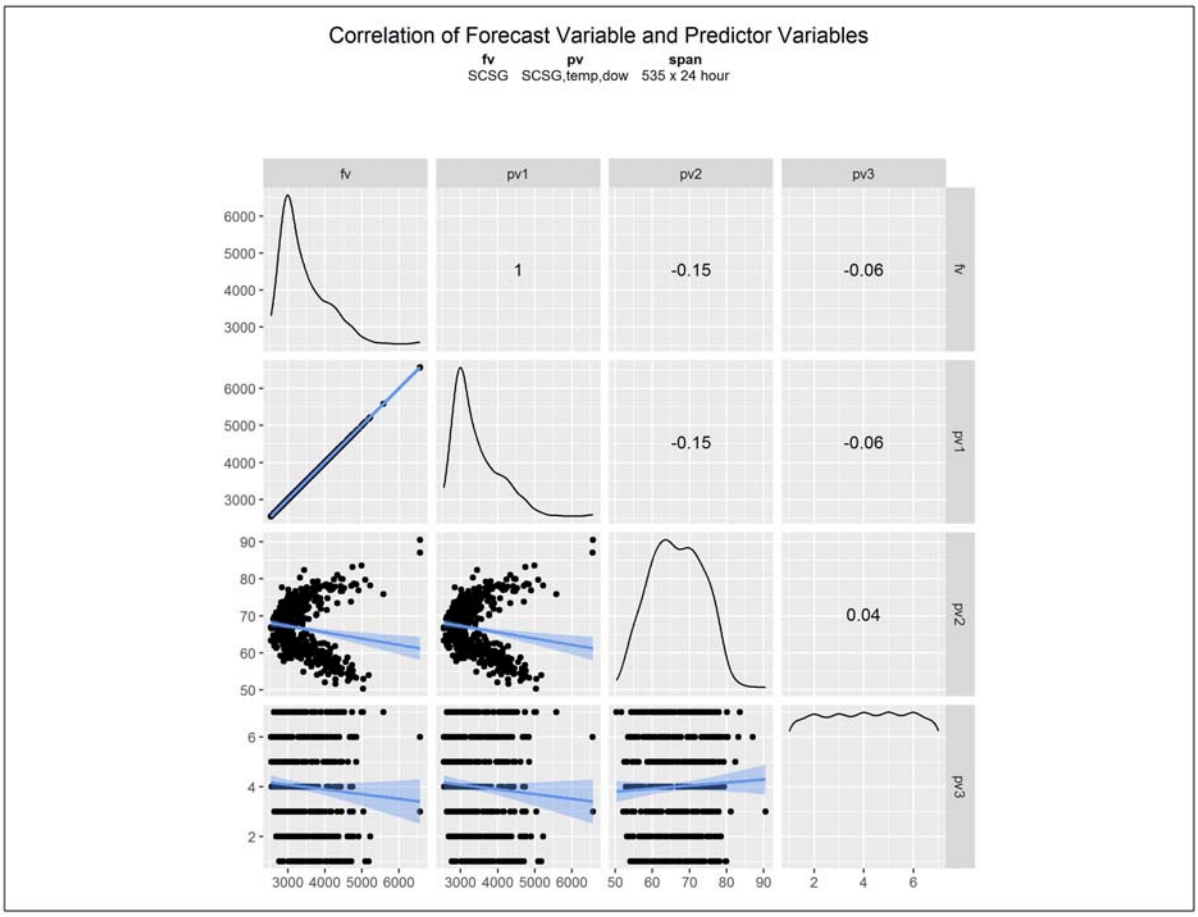


Figure 6-4: Correlations of series, Australia, for week-ahead forecasts. Reference series is aggregated 223-household electricity usage. Predictor series are aggregate electricity usage, temperature, and day of week (integer coded).

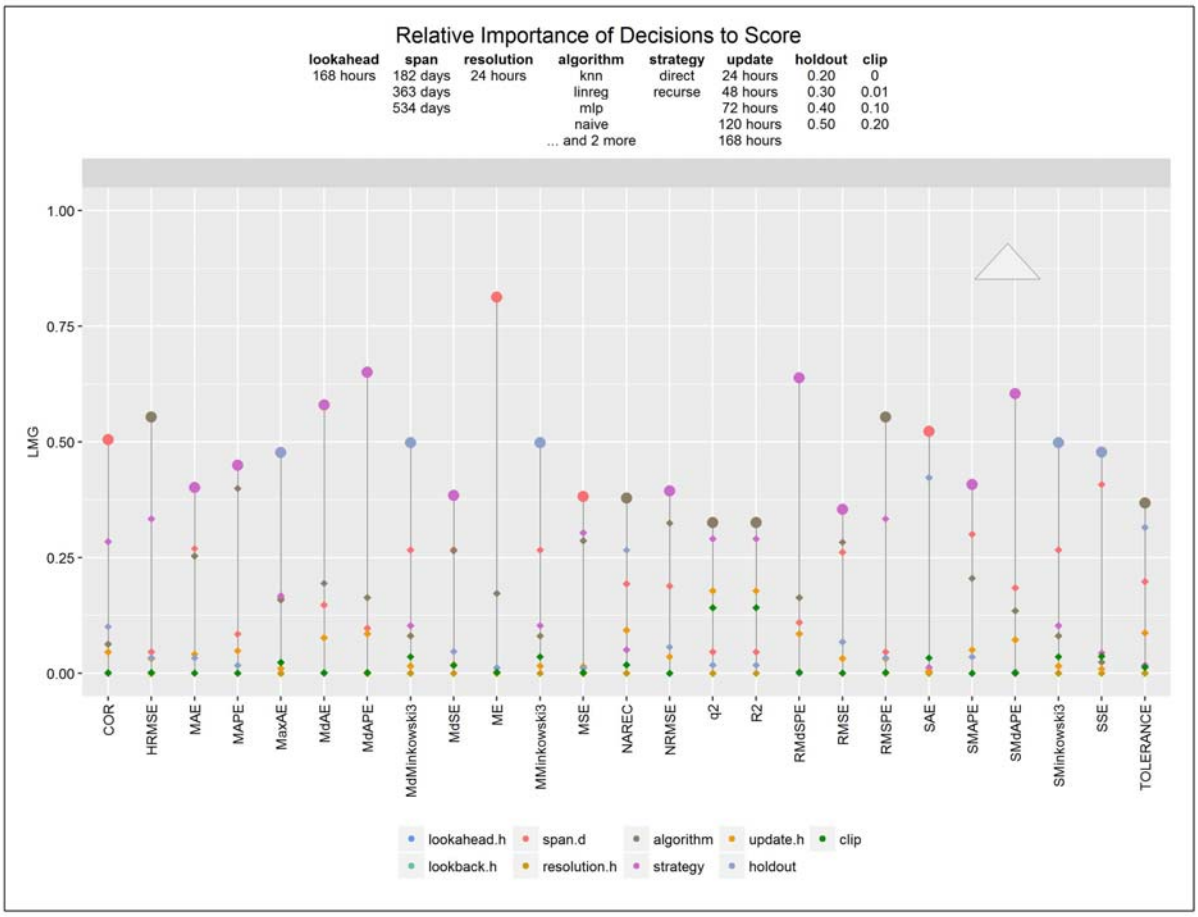


Figure 6-5: Relative importance of decision to metric score variation, across metrics, Australia, week-ahead forecasts. 2,880 techniques and forecasts. LMG score for each decision is presented along y-axis. Metrics are arranged along x-axis. *Colors* are decision. *Large point* is highest relative importance.

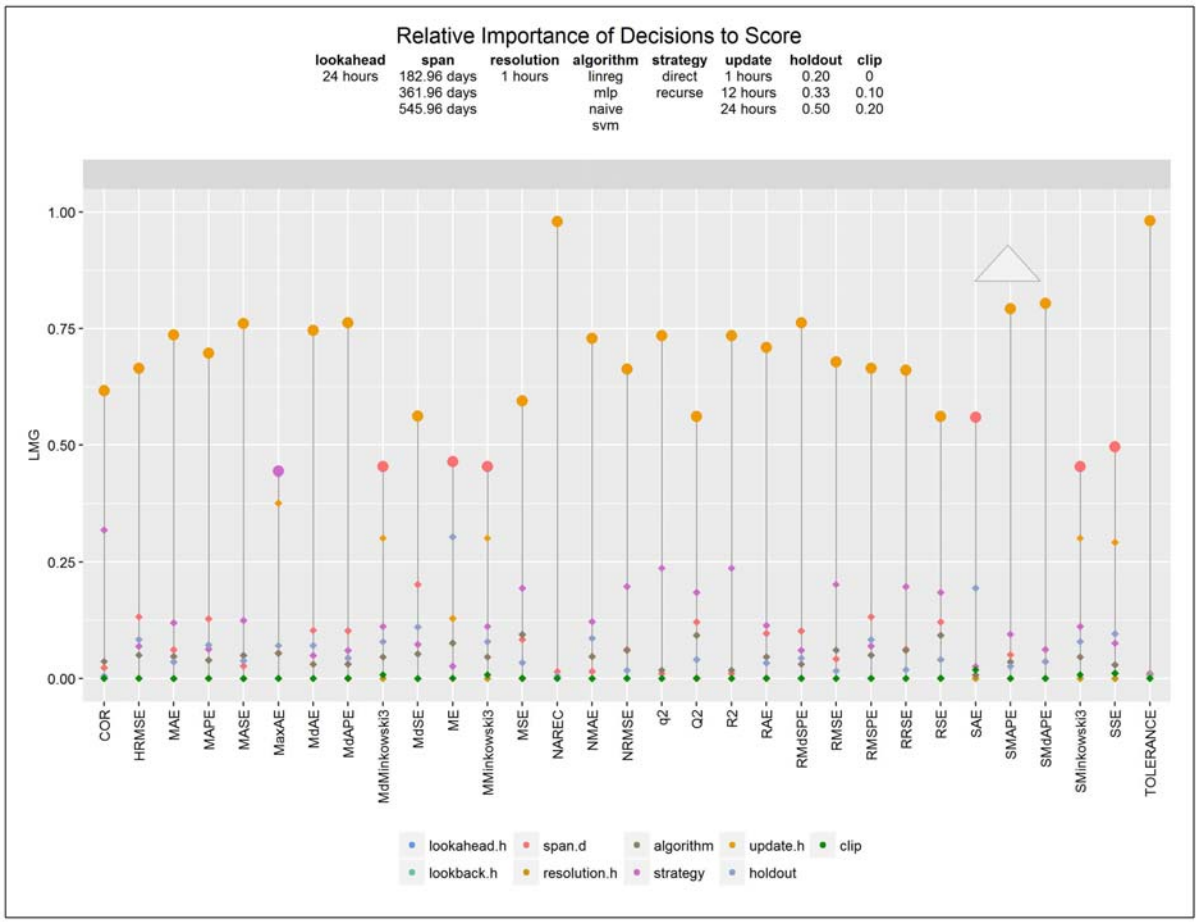


Figure 6-6: Relative importance of decision to metric score variation, across metrics, Australia, day-ahead forecasts. 648 techniques and forecasts. LMG score for each decision is presented along y-axis. Metrics are arranged along x-axis. *Colors* are decision. *Large point* is highest relative importance.

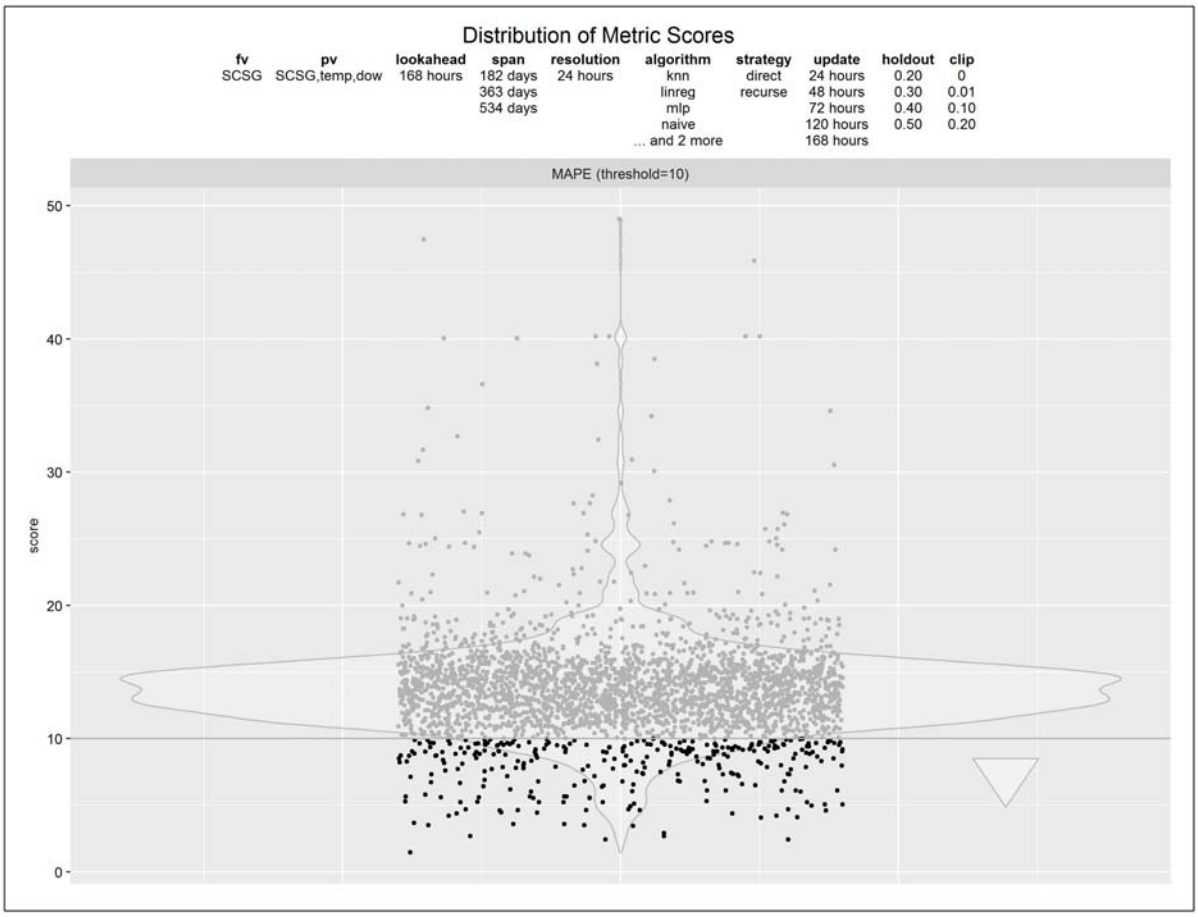


Figure 6-7: Distribution of metric scores, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Metric is MAPE. *Black* is metric score for a qualified technique. *Gray* is metric score for an unqualified technique.

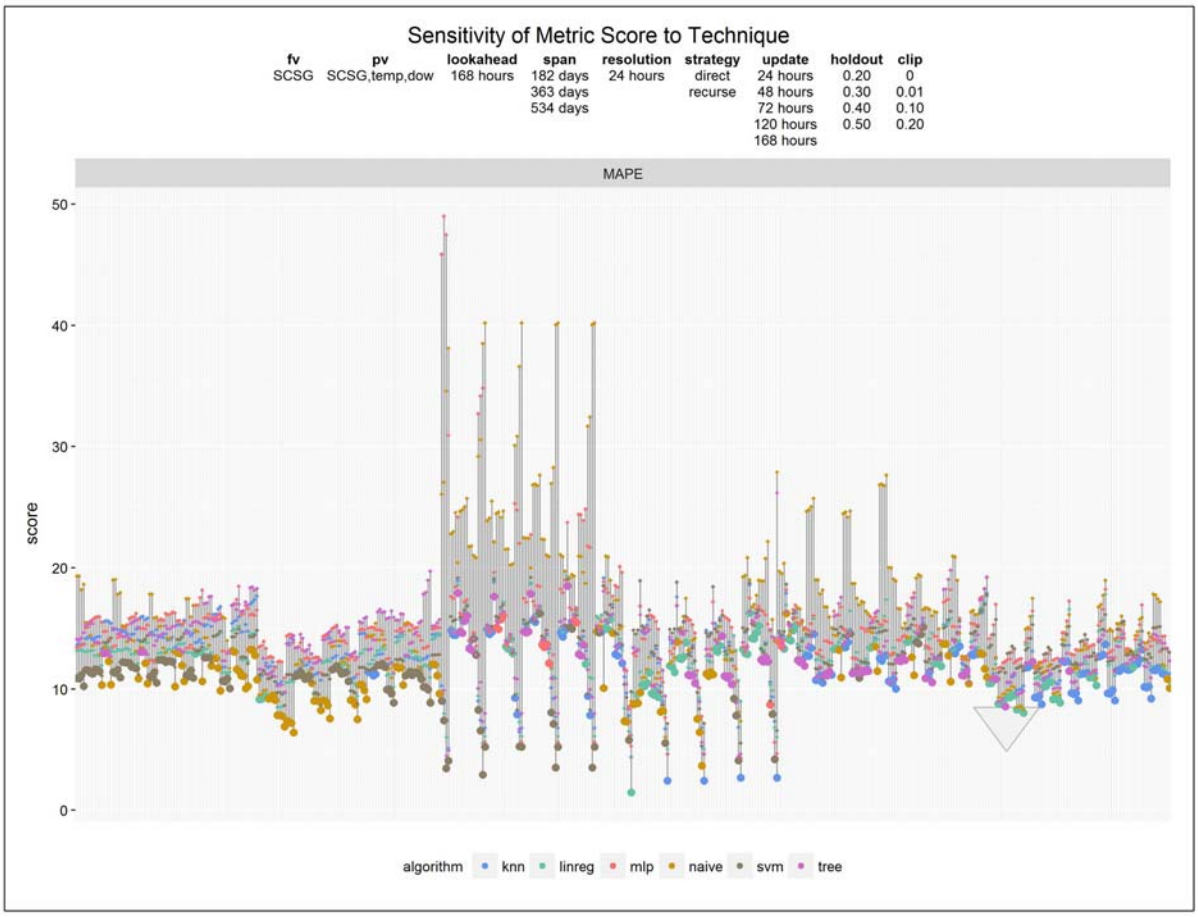


Figure 6-8: Distribution of metric scores across techniques, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each vertical bar represents a family of techniques that differ from each other only by algorithm class. Distances along the vertical bars represent metric scores for specific techniques. Metric is MAPE. *Colors* are algorithm class. *Large point* is metric score of best technique within a family of techniques.

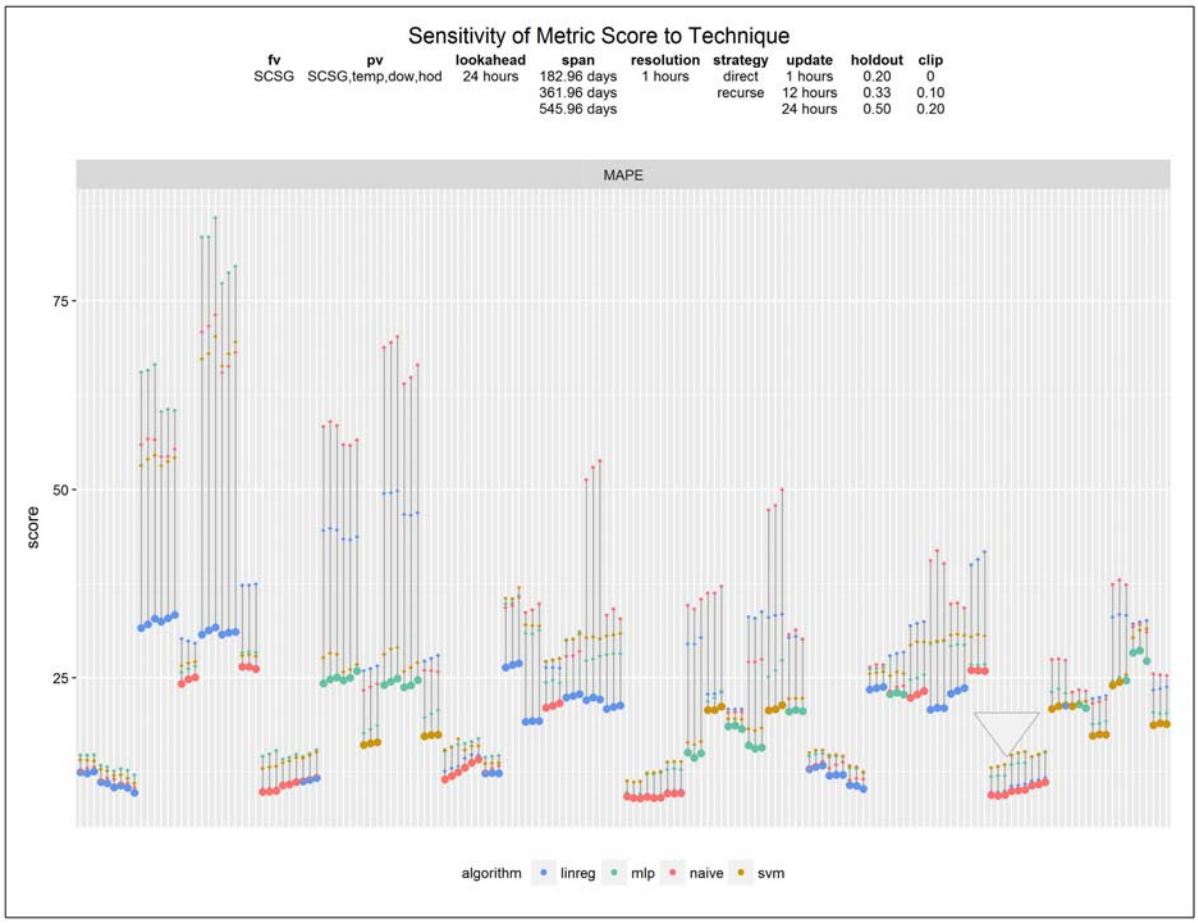


Figure 6-9: Distribution of metric scores across techniques, Australia, day-ahead forecasts. 648 techniques and forecasts. Each vertical bar represents a family of techniques that differ from each other only by algorithm class. Distances along the vertical bars represent metric scores for specific techniques. Metric is MAPE. *Colors* are algorithm class. *Large point* is metric score of best technique within a family of techniques.

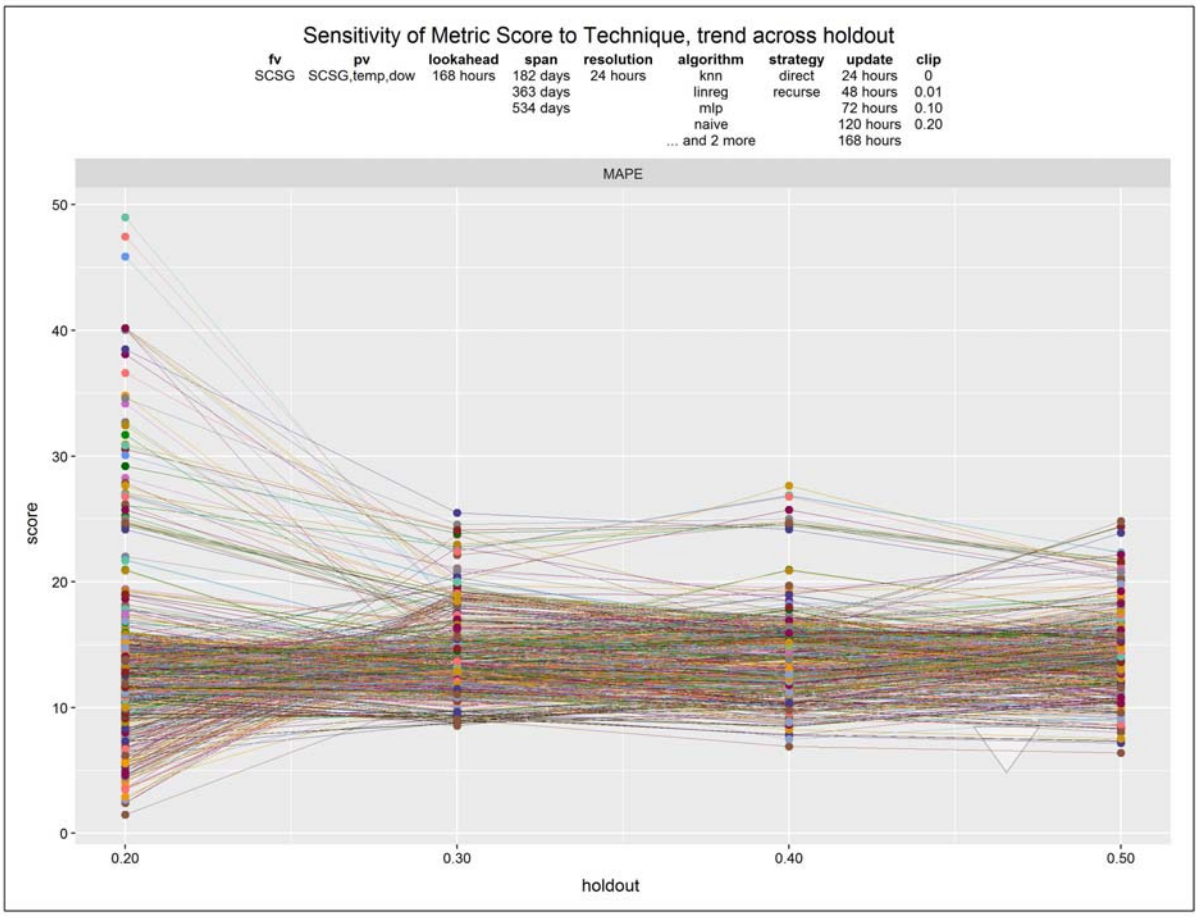


Figure 6-10: Metric score vs. holdout decision option, split by technique family, for several metrics, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each trend line represents a family of techniques that differ from each other only by holdout decision option. Holdouts are arranged sequentially along the x-axis. Metric is MAPE. *Colors* are technique family.

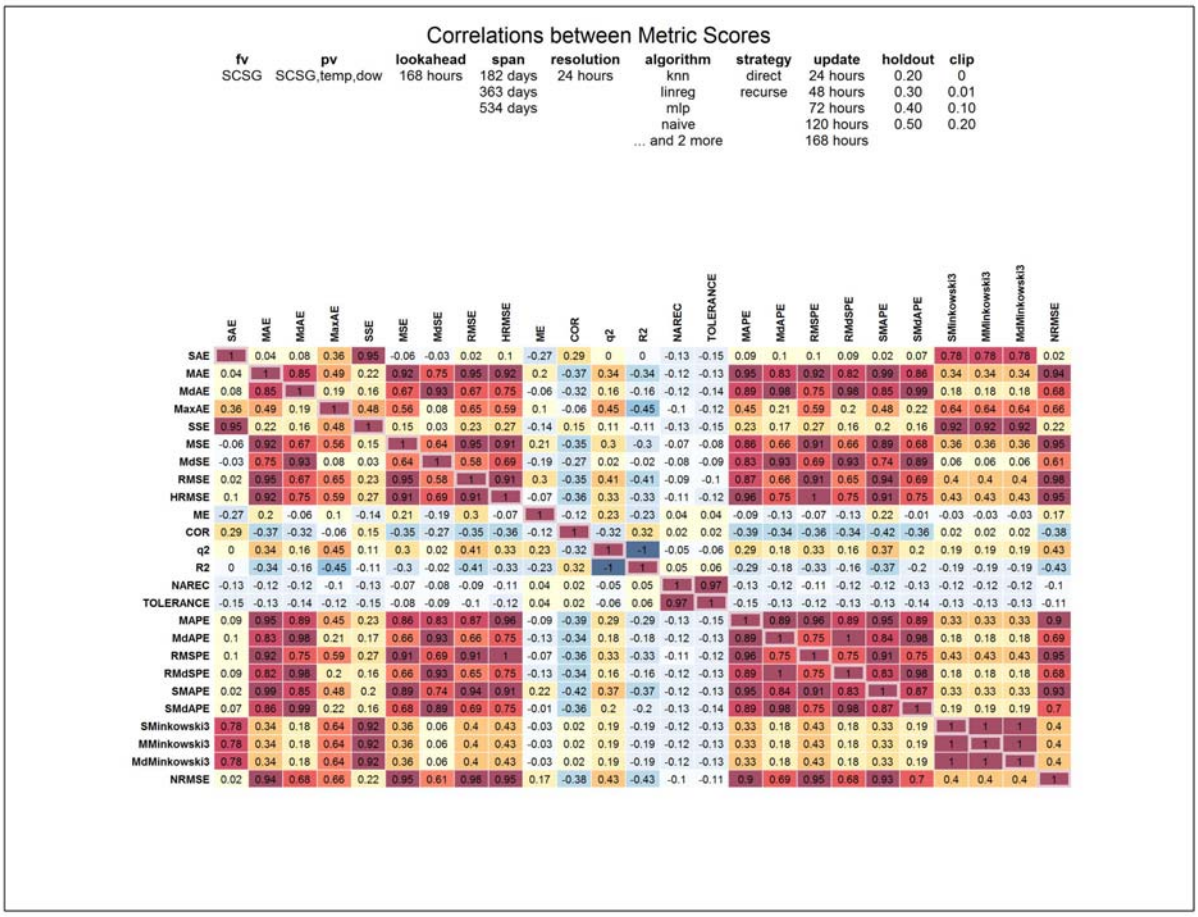


Figure 6-11: Correlations of scores per one metric-to-scores per another metric, across metric pairs, Australia, week-ahead forecasts. 2,880 techniques and forecasts. For each cell, all techniques are scored per two metrics, and the sequence of scores per the first metric are correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation.

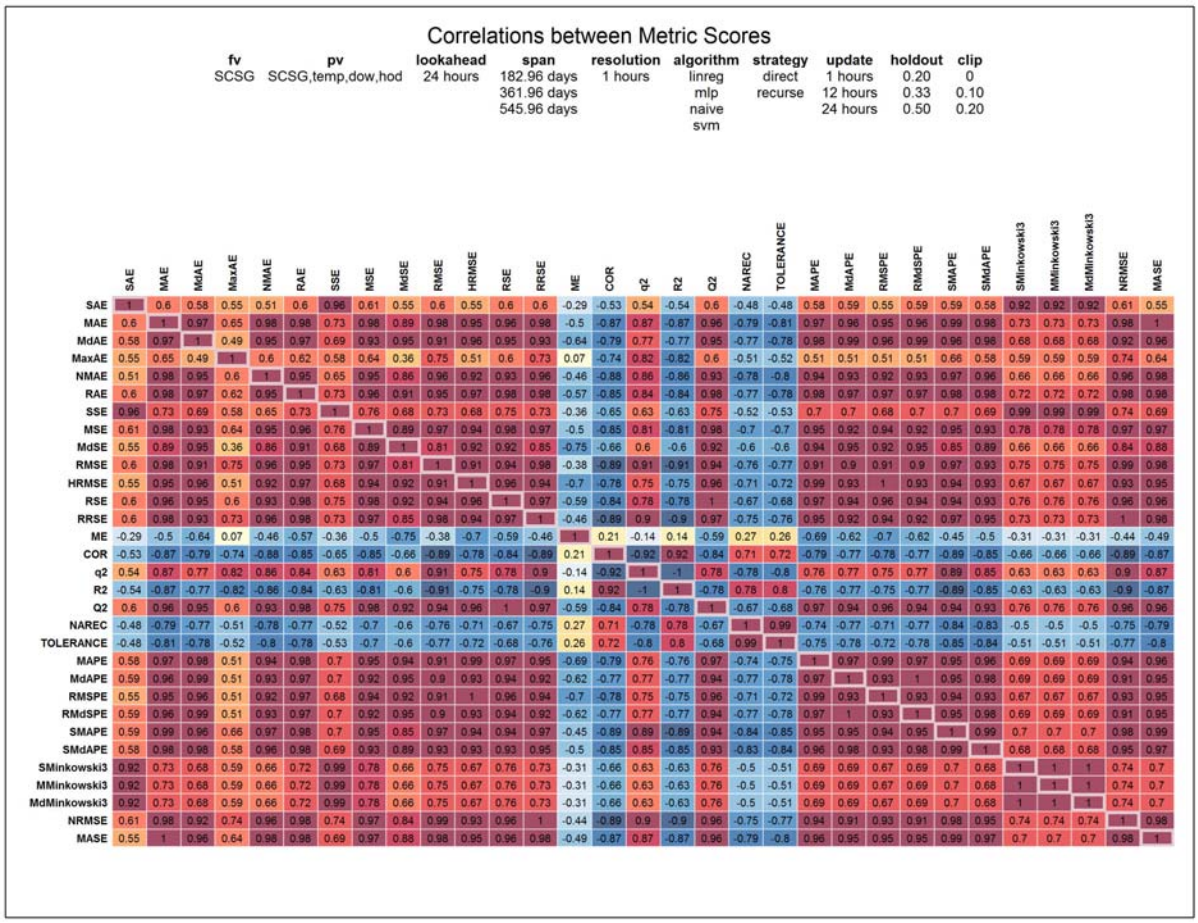


Figure 6-12: Correlations of scores per one metric-to-scores per another metric, across metric pairs, Australia, day-ahead forecasts. 648 techniques and forecasts. For each cell, all techniques are scored per two metrics, and the sequence of scores per the first metric are correlated with the sequence of scores per the second metric. *Red* is positive correlation. *Blue* is negative correlation. *Dark* is strong absolute correlation. *Light* is weak absolute correlation.

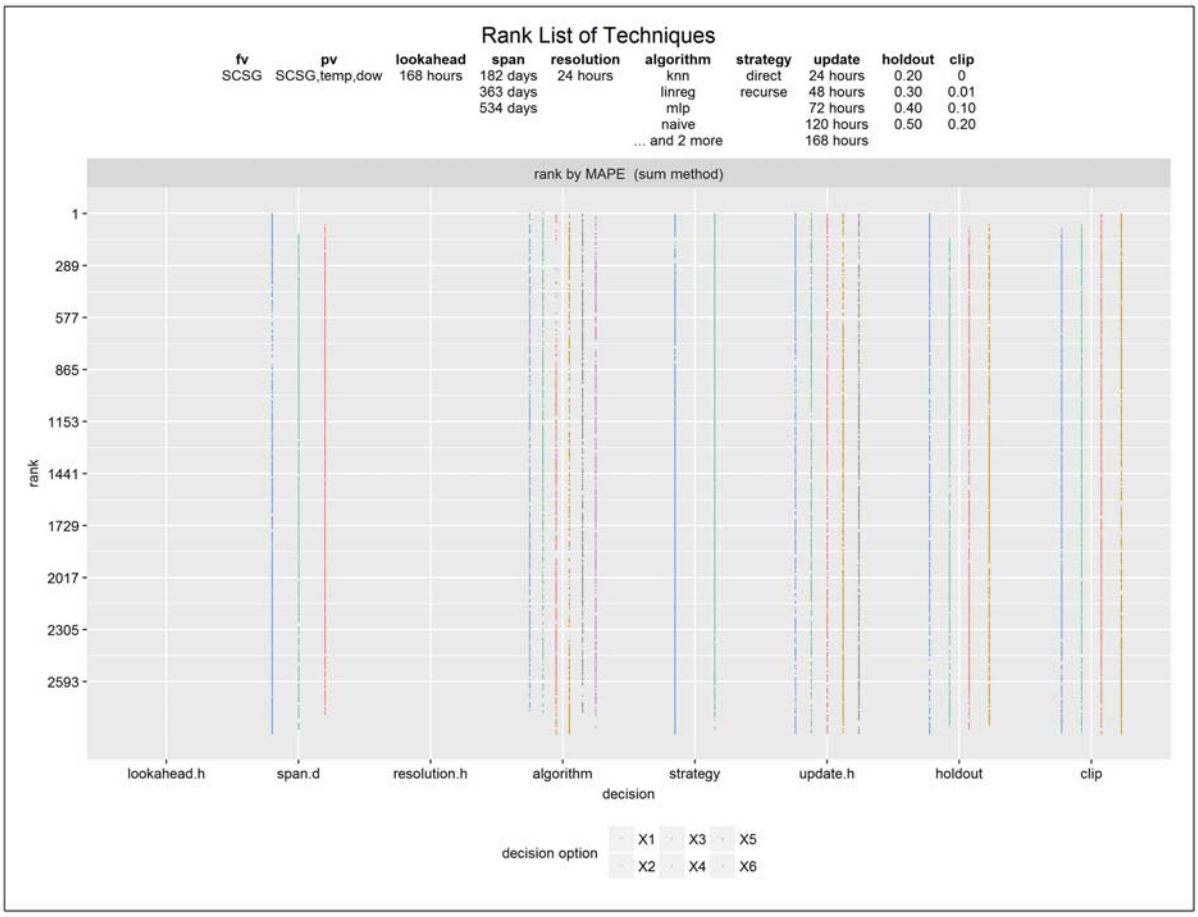


Figure 6-13: List of techniques, rank ordered by performance per MAPE score, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

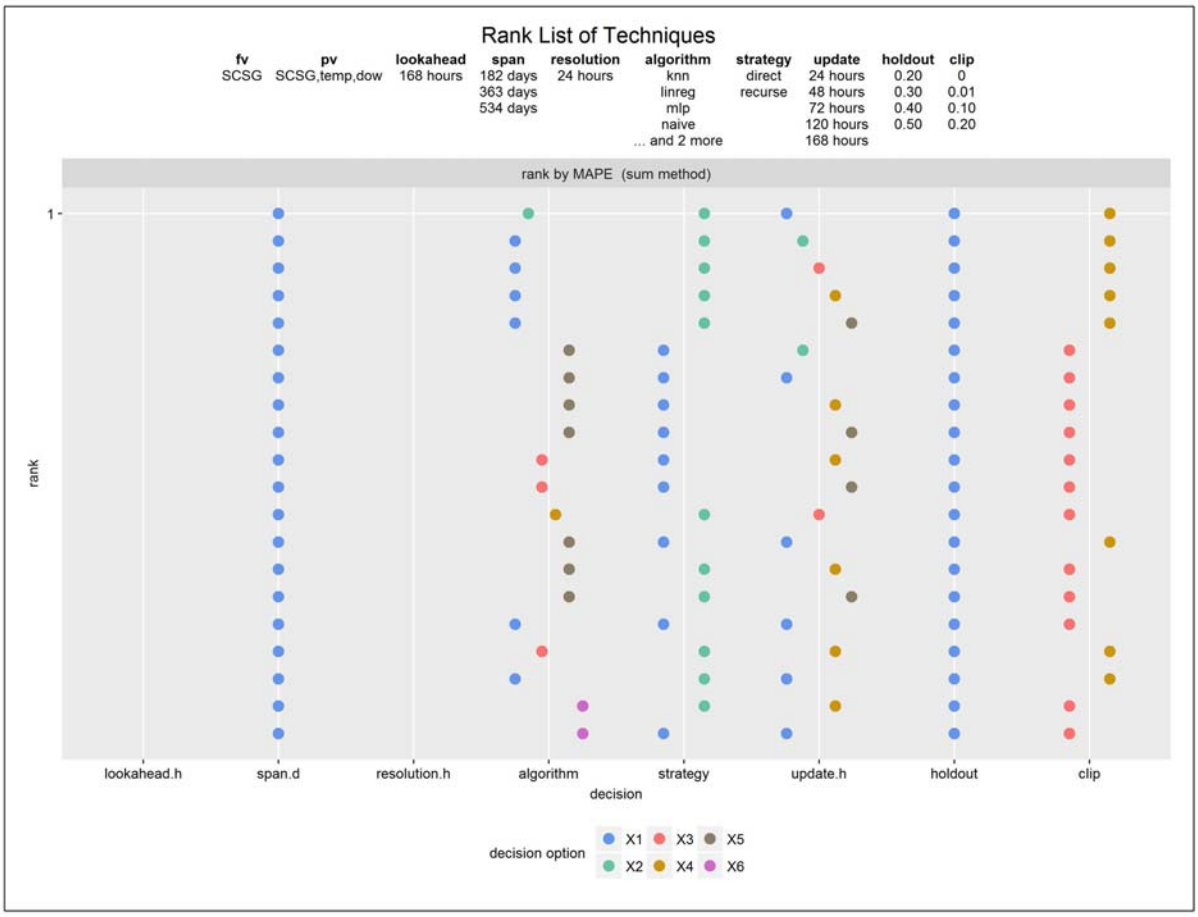


Figure 6-14: List of best techniques, rank ordered by performance per MAPE score, Australia, week-ahead forecasts. 20 best of 2,880 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

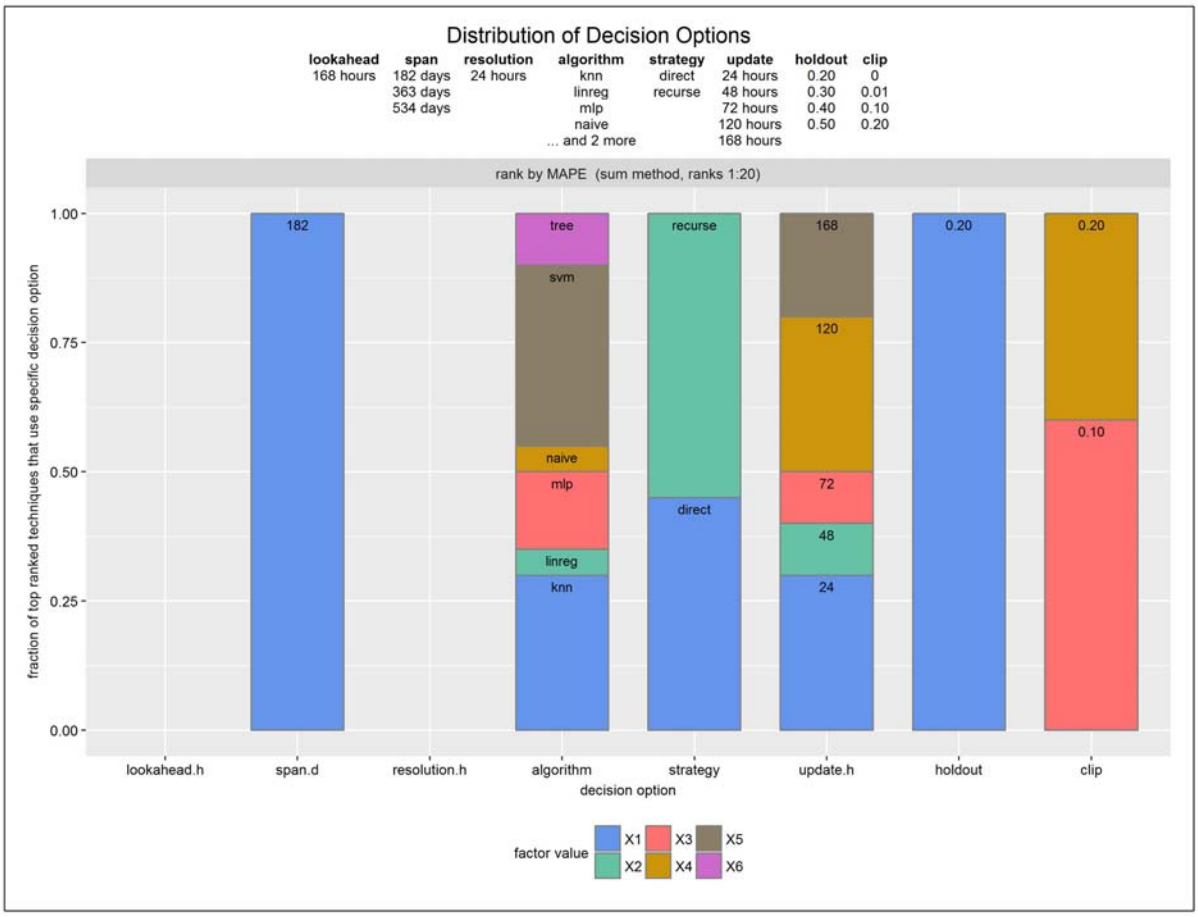


Figure 6-15: Distribution of decision options in best techniques, Australia, week-ahead forecasts. 20 best of 2,880 techniques and forecasts. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

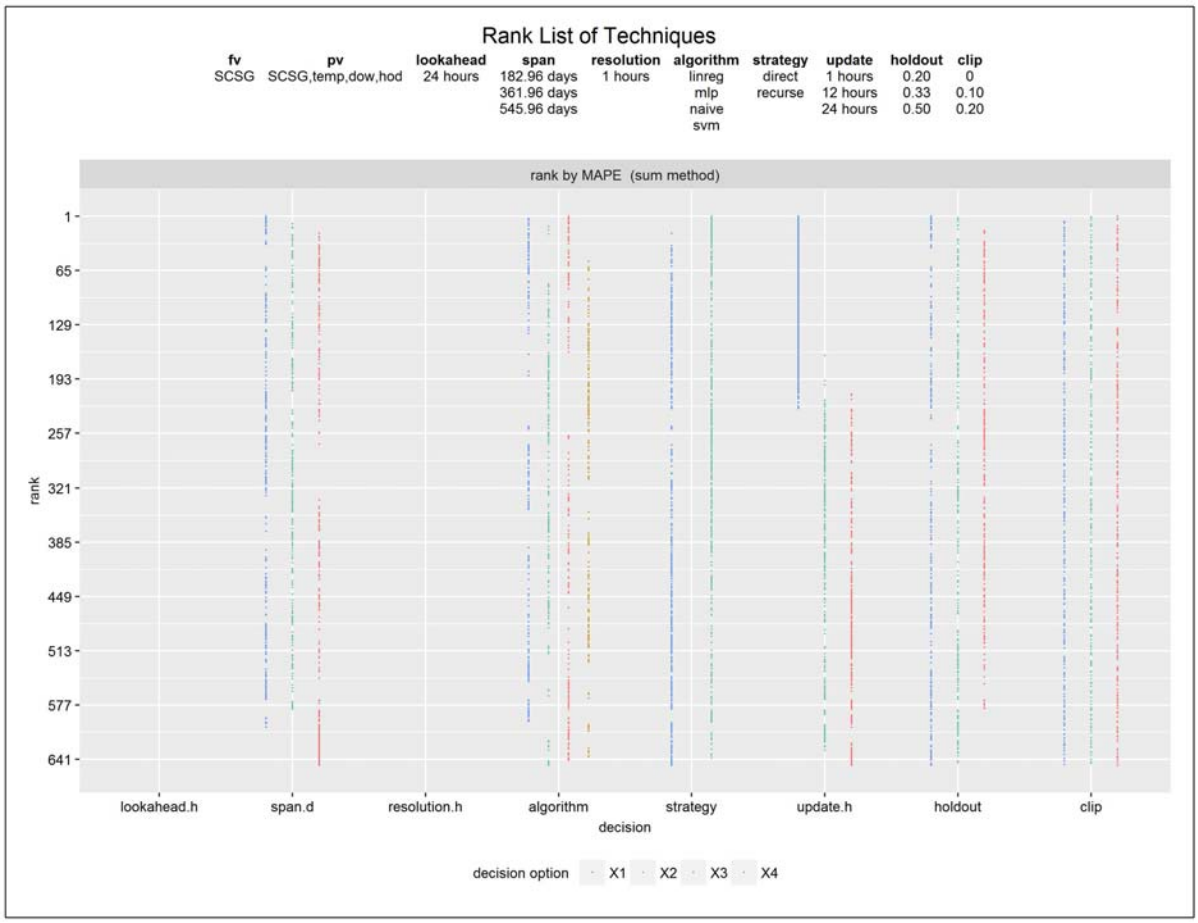


Figure 6-16: List of techniques, rank ordered by performance per MAPE score, Australia, day-ahead forecasts. 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

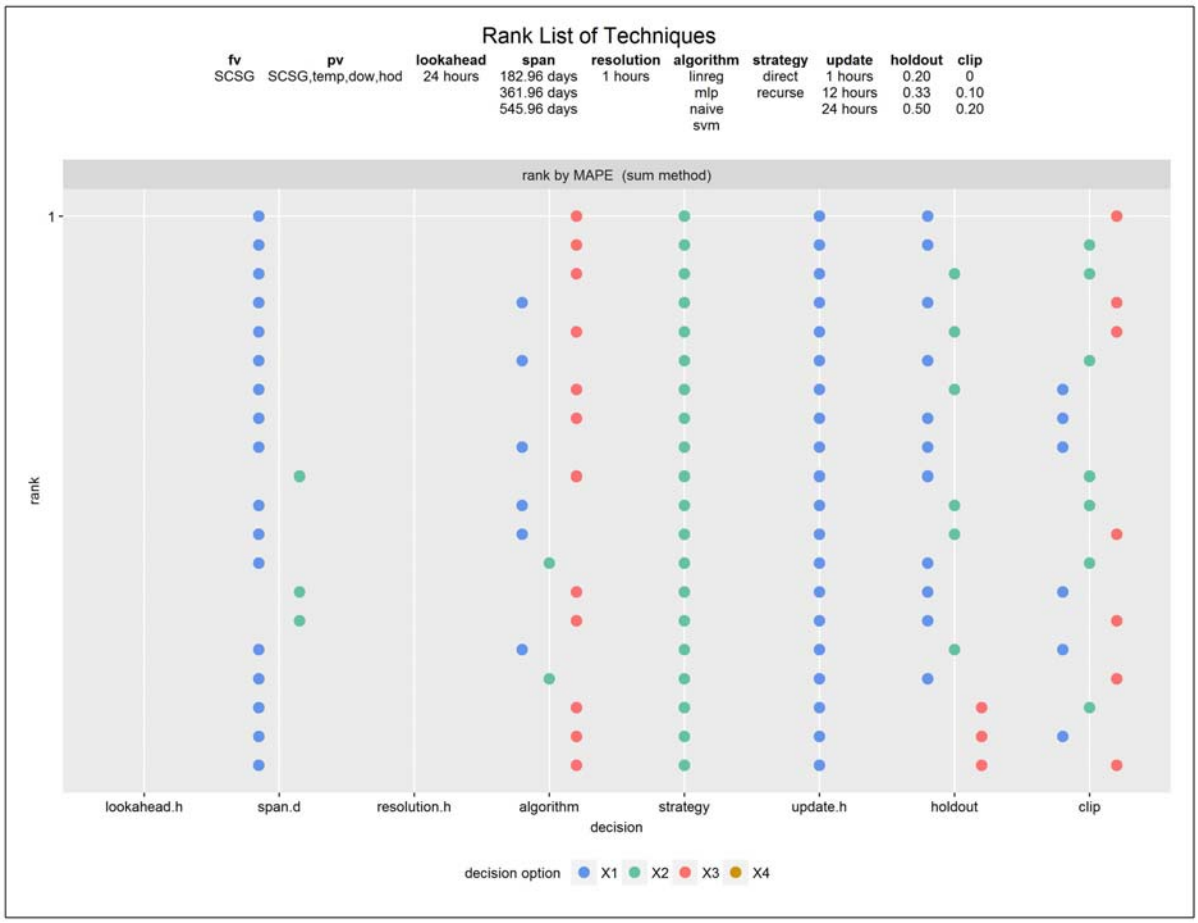


Figure 6-17: List of best techniques, rank ordered by performance per MAPE score, Australia, day-ahead forecasts. 20 best of 648 techniques and forecasts. Rank 1 means best performing technique. Decisions are arranged as major columns along the x-axis. Decision options are arranged as minor columns within a major column along the x-axis. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

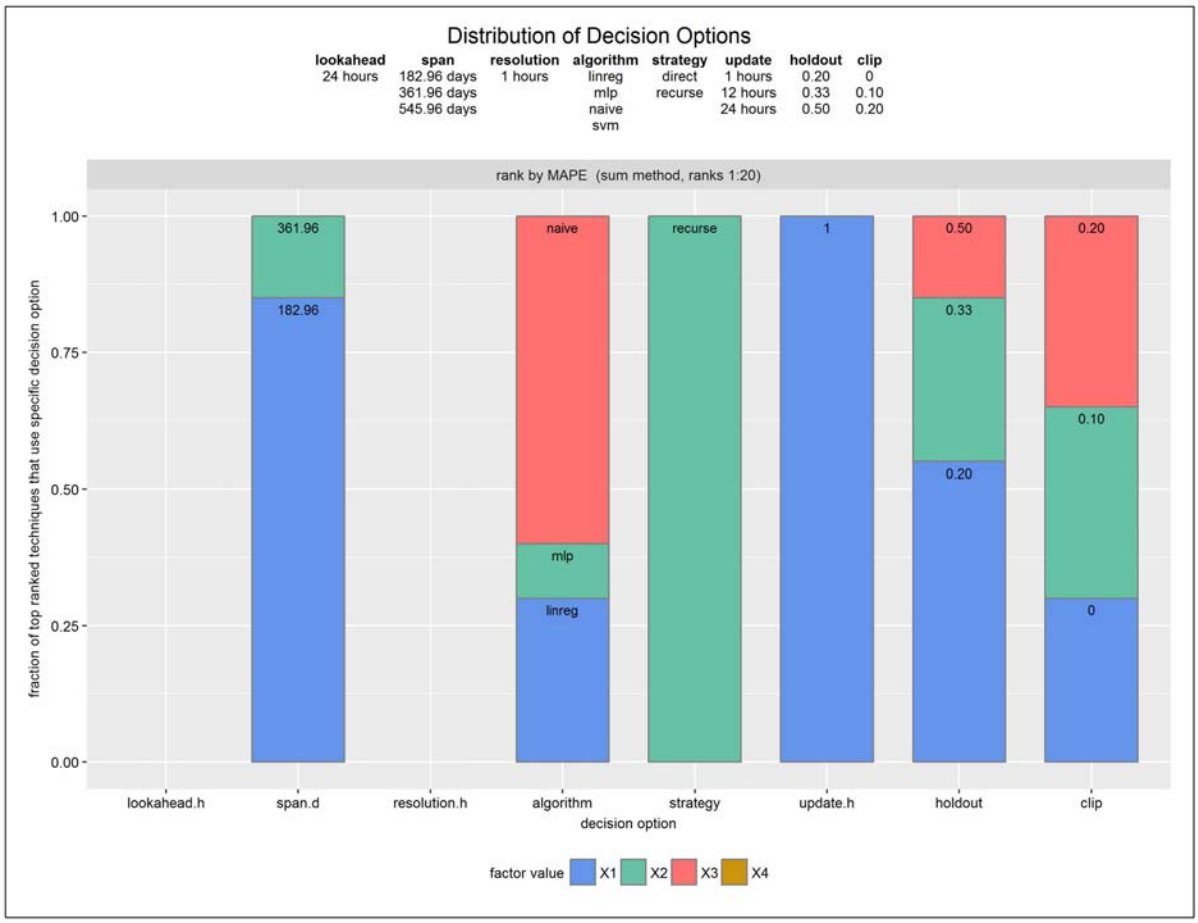


Figure 6-18: Distribution of decision options in best techniques, Australia, day-ahead forecasts. 20 best of 648 techniques and forecasts. *Colors* are the decision option – X1 means the first decision option, X2 means the second decision option, etc. Grayed major columns indicate decisions that are not varied by forecasting practitioner.

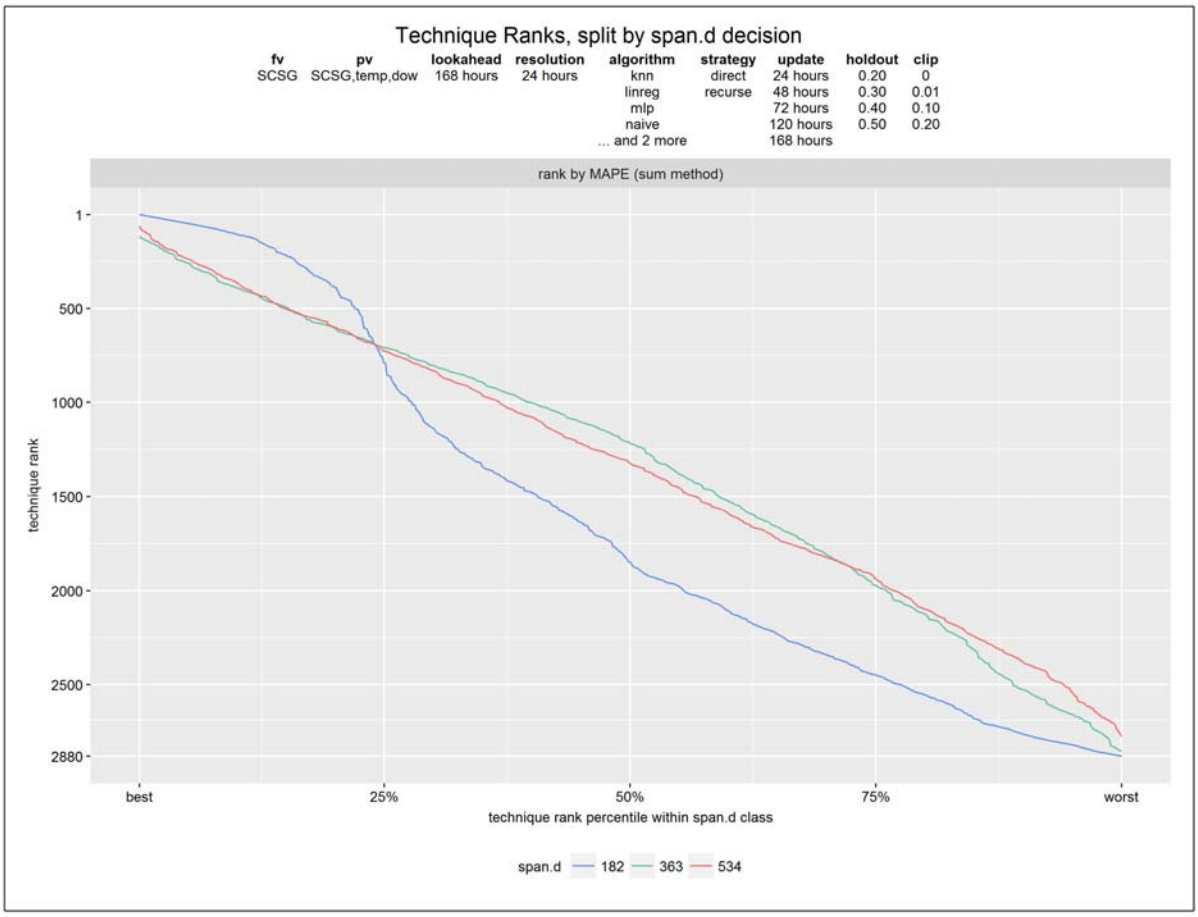


Figure 6-19: Technique rank trend, split by span, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the span decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the span decision option.

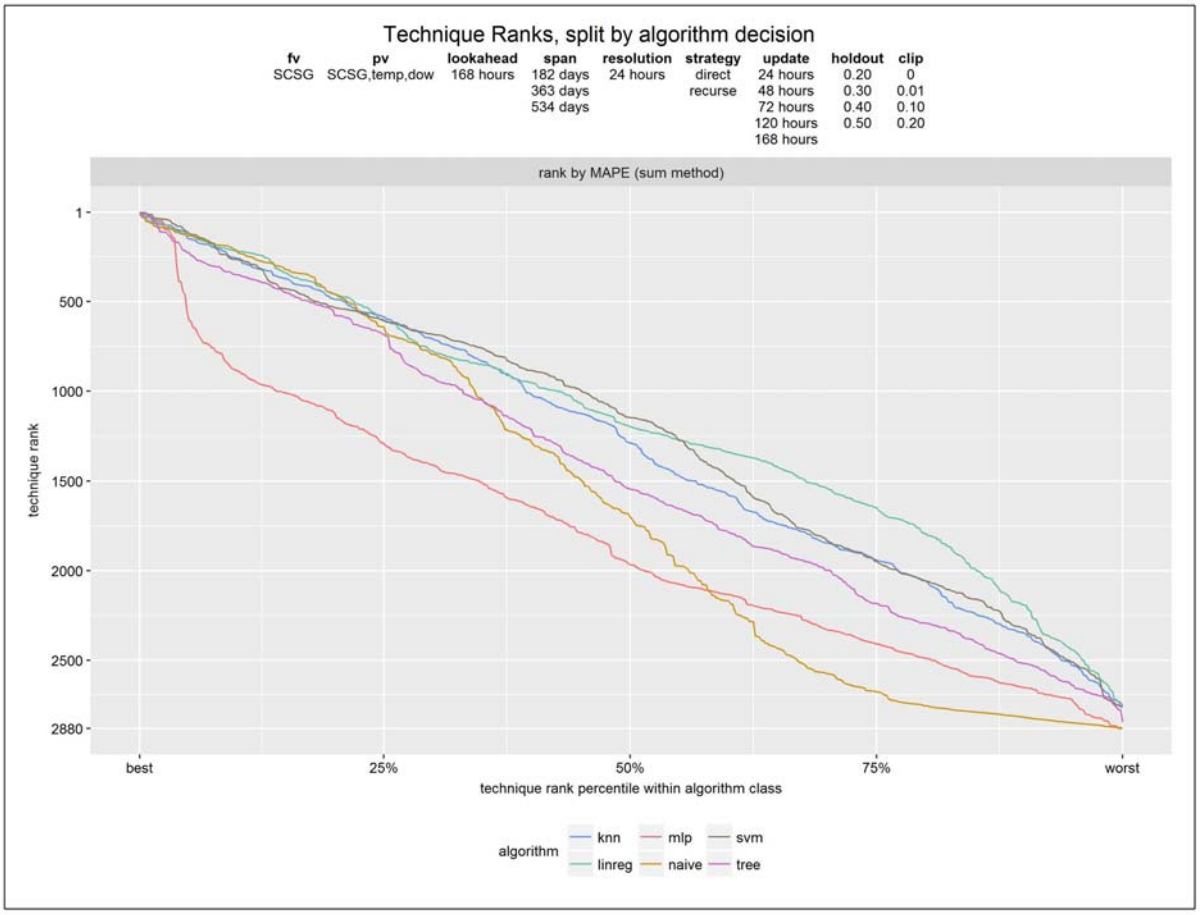


Figure 6-20: Technique rank trend, split by algorithm class, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

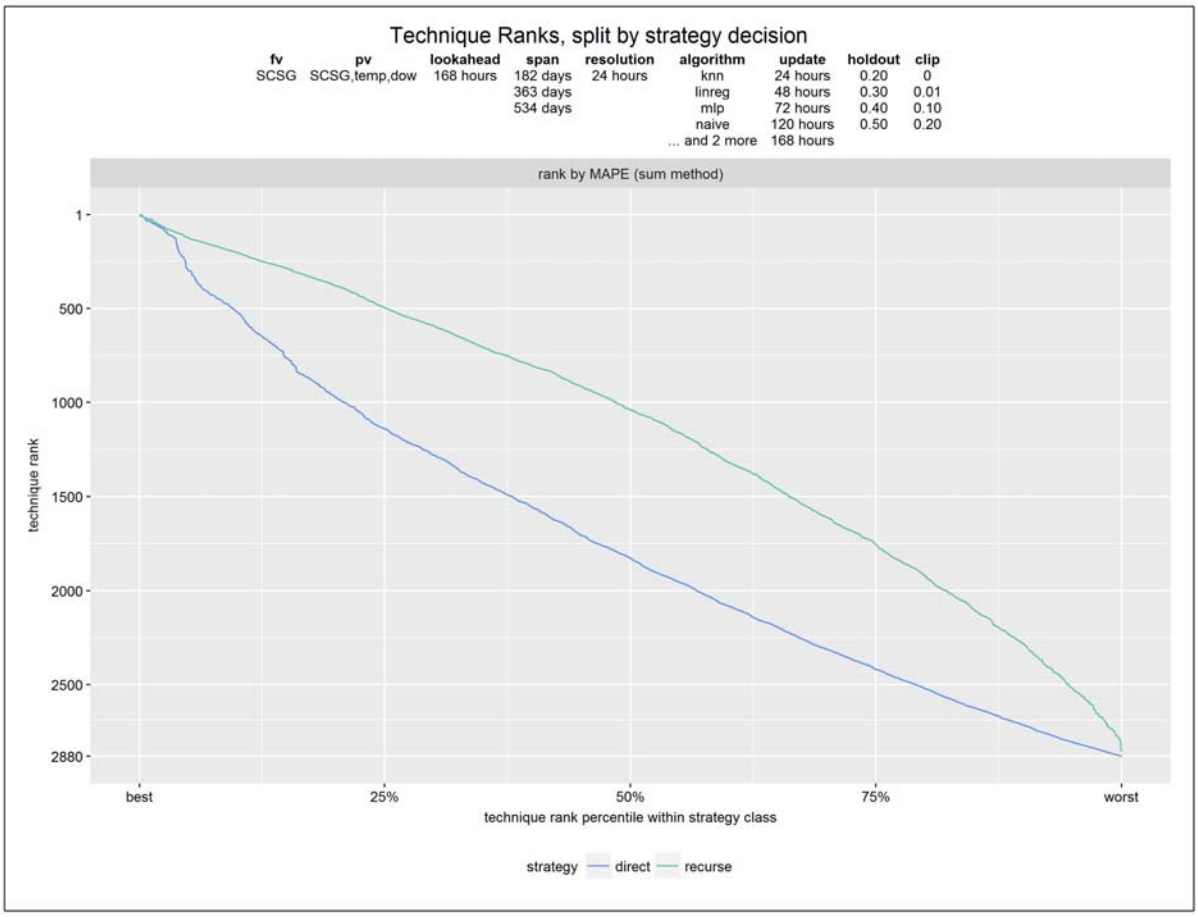


Figure 6-21: Technique rank trend, split by extension rule, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the extension rule decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the extension rule decision option.

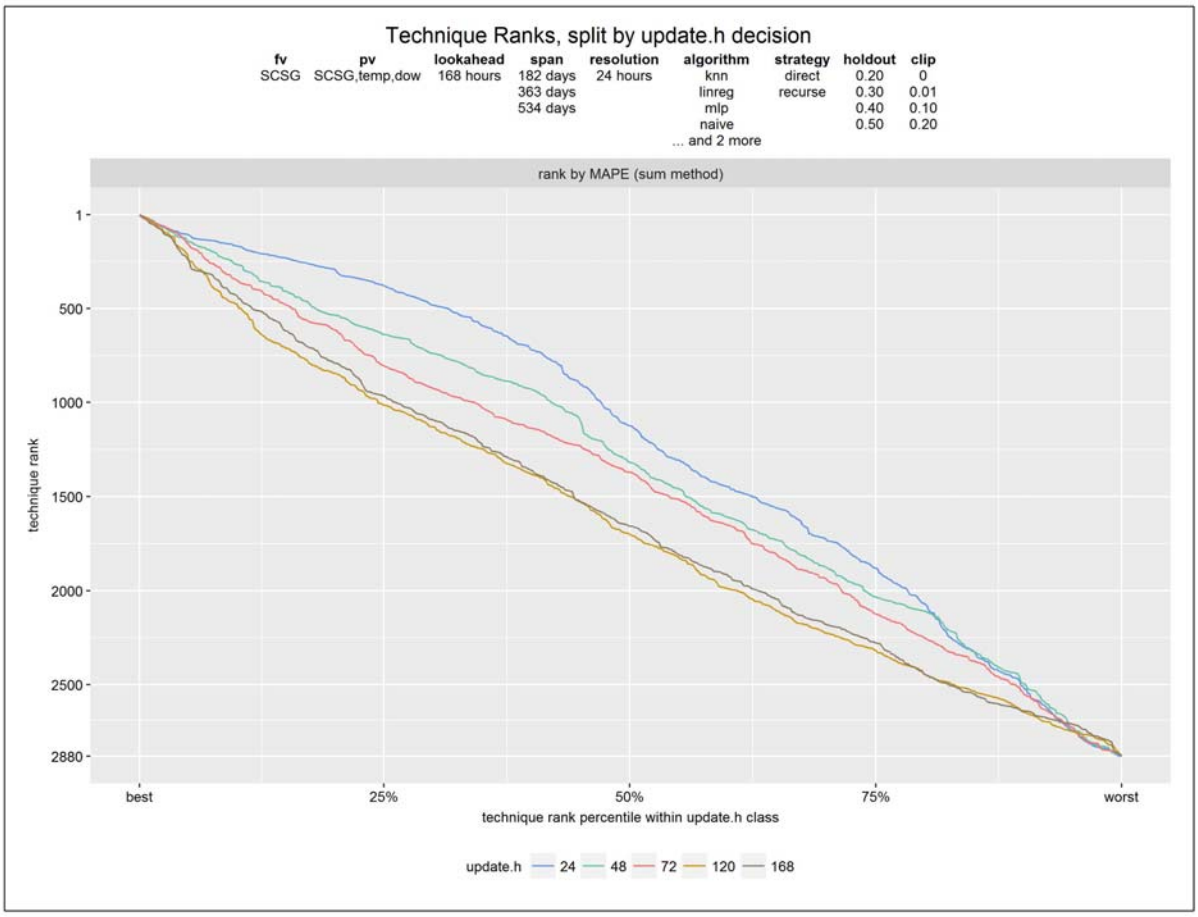


Figure 6-22: Technique rank trend, split by update cycle, Australia, week-ahead forecast. 2,880 techniques and forecasts. Each curve represents a family of techniques with the update cycle decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the update cycle decision option.

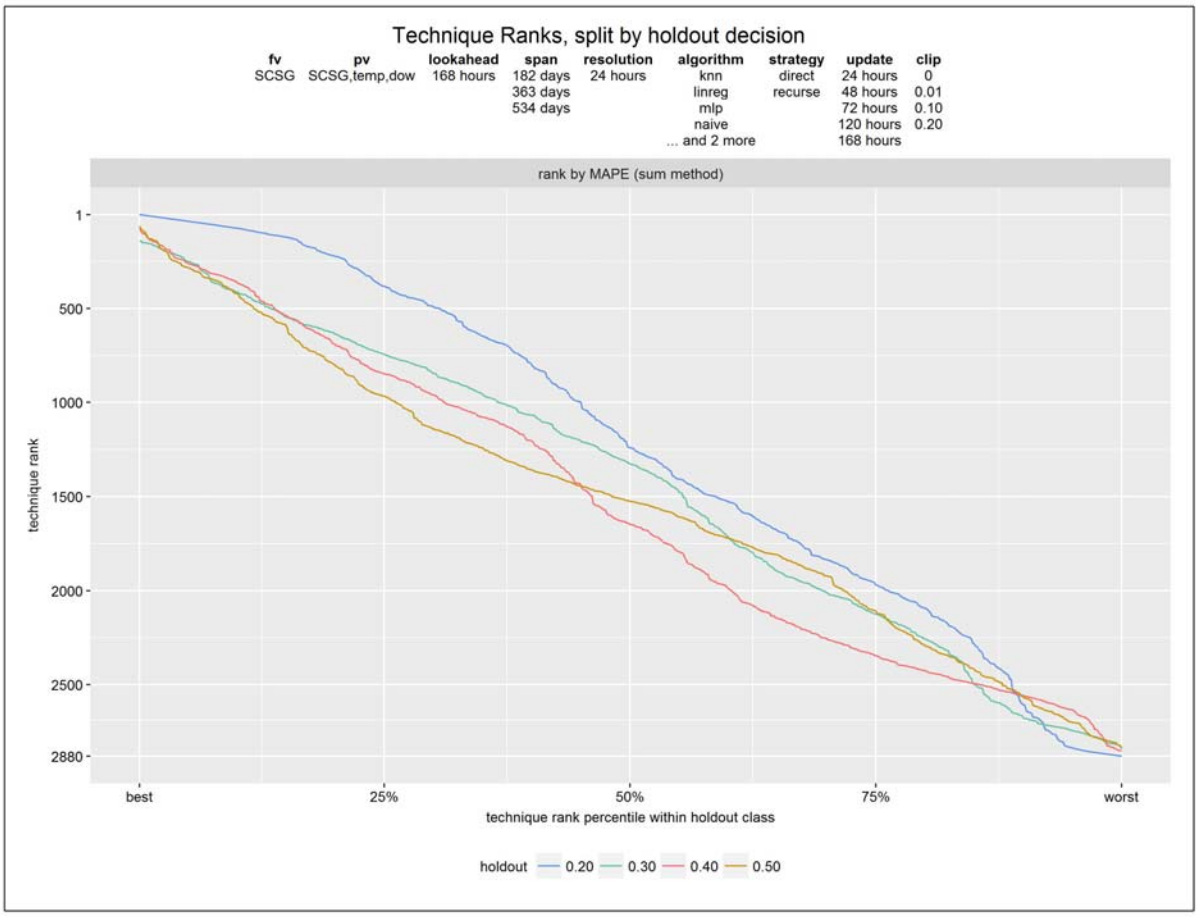


Figure 6-23: Technique rank trend, split by holdout, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the holdout decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the holdout decision option.

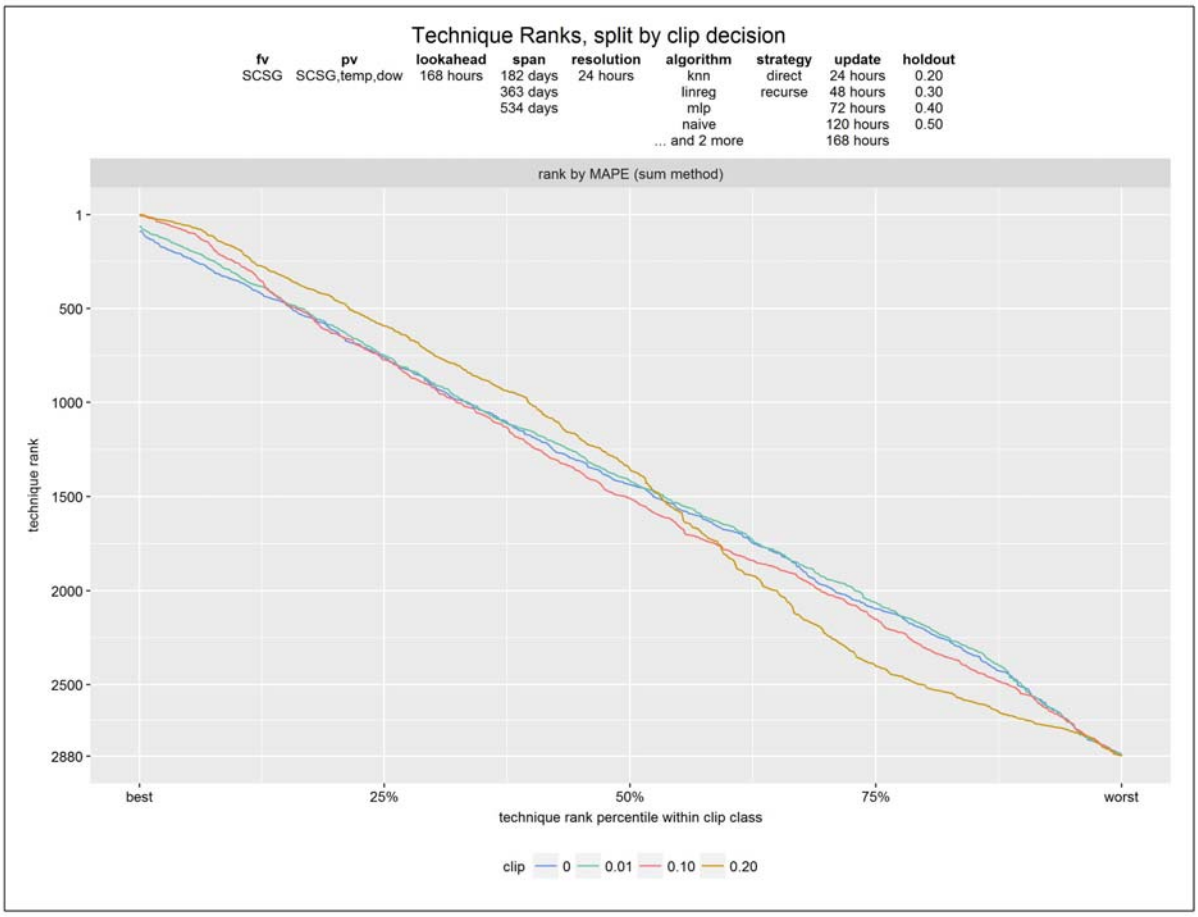


Figure 6-24: Technique rank trend, split by clip, Australia, week-ahead forecasts. 2,880 techniques and forecasts. Each curve represents a family of techniques with the clip decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the clip decision option.

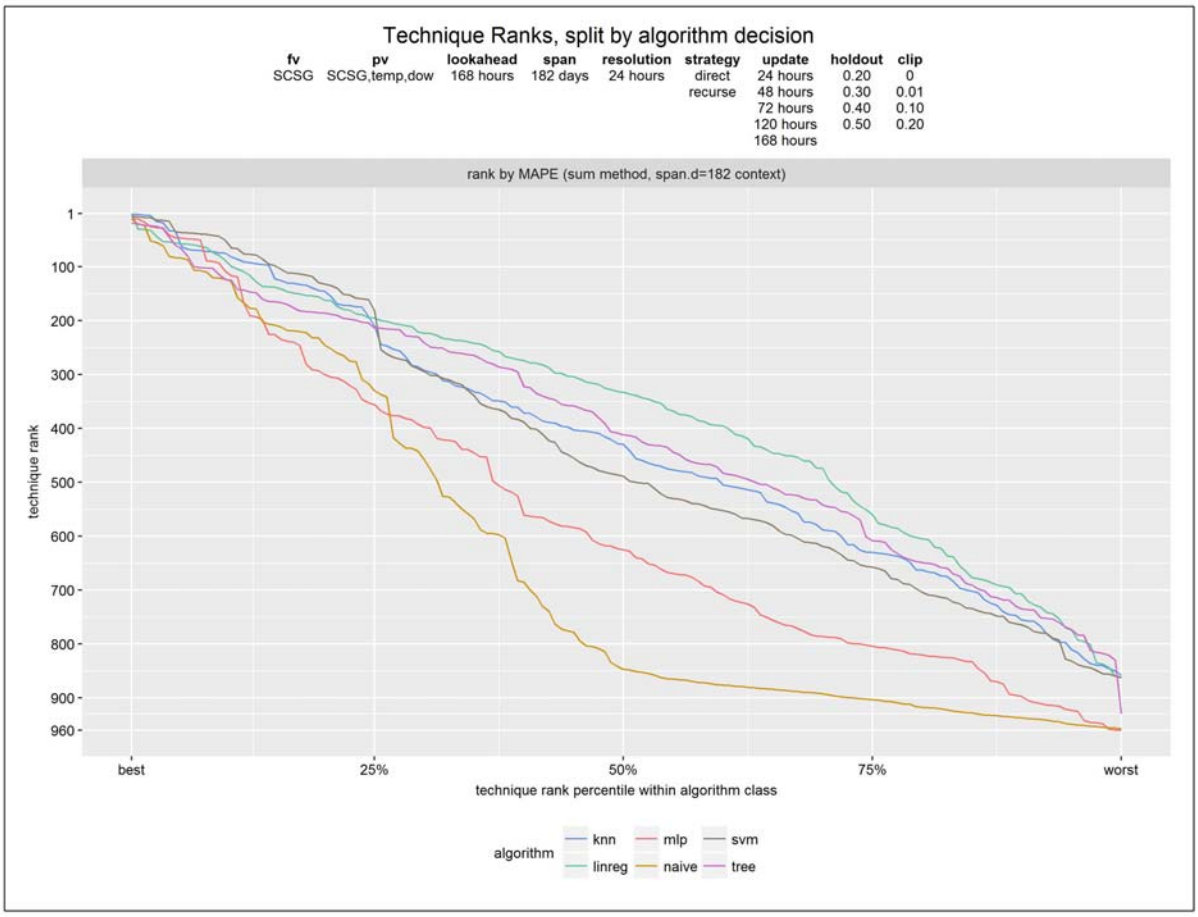


Figure 6-25: Technique rank trend at short span, split by algorithm class, Australia, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 182-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

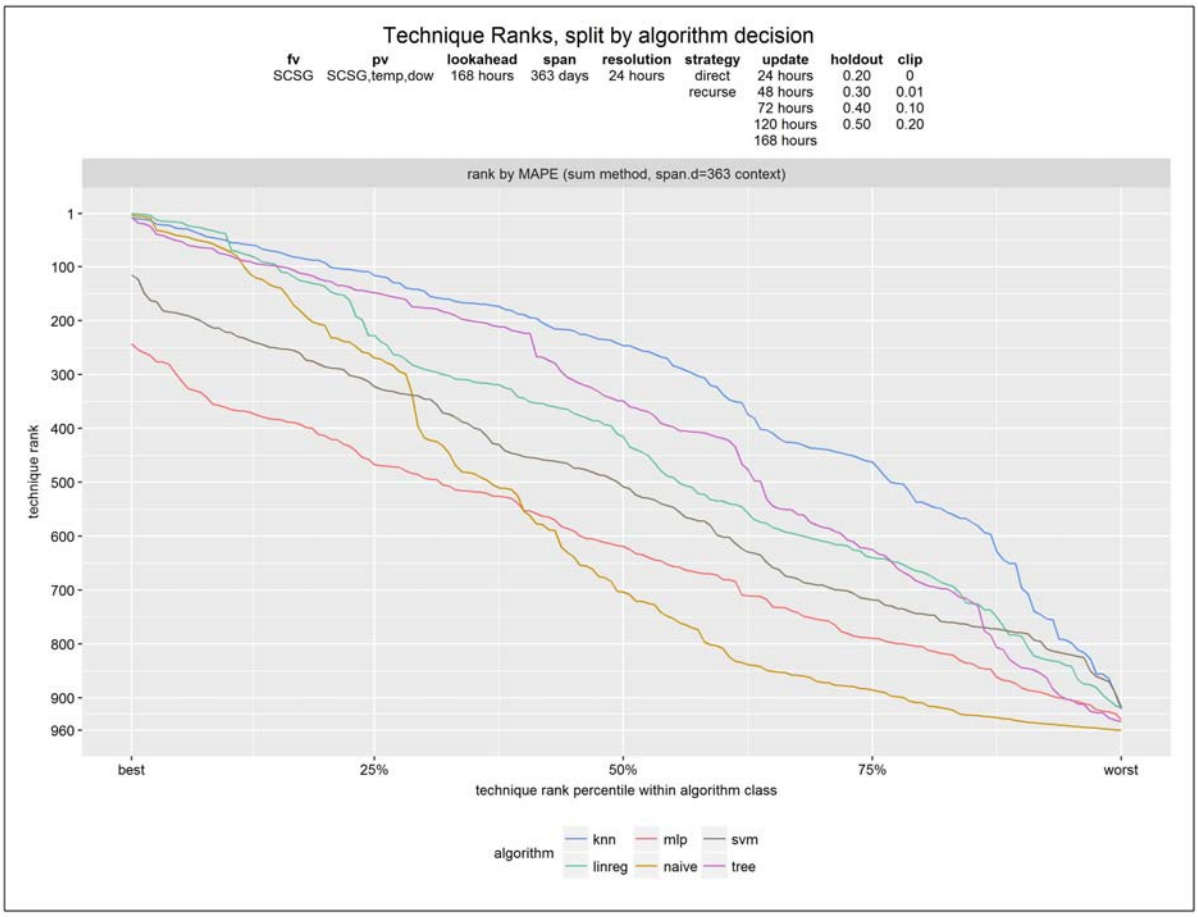


Figure 6-26: Technique rank trend at medium span, split by algorithm class, Australia, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 363-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

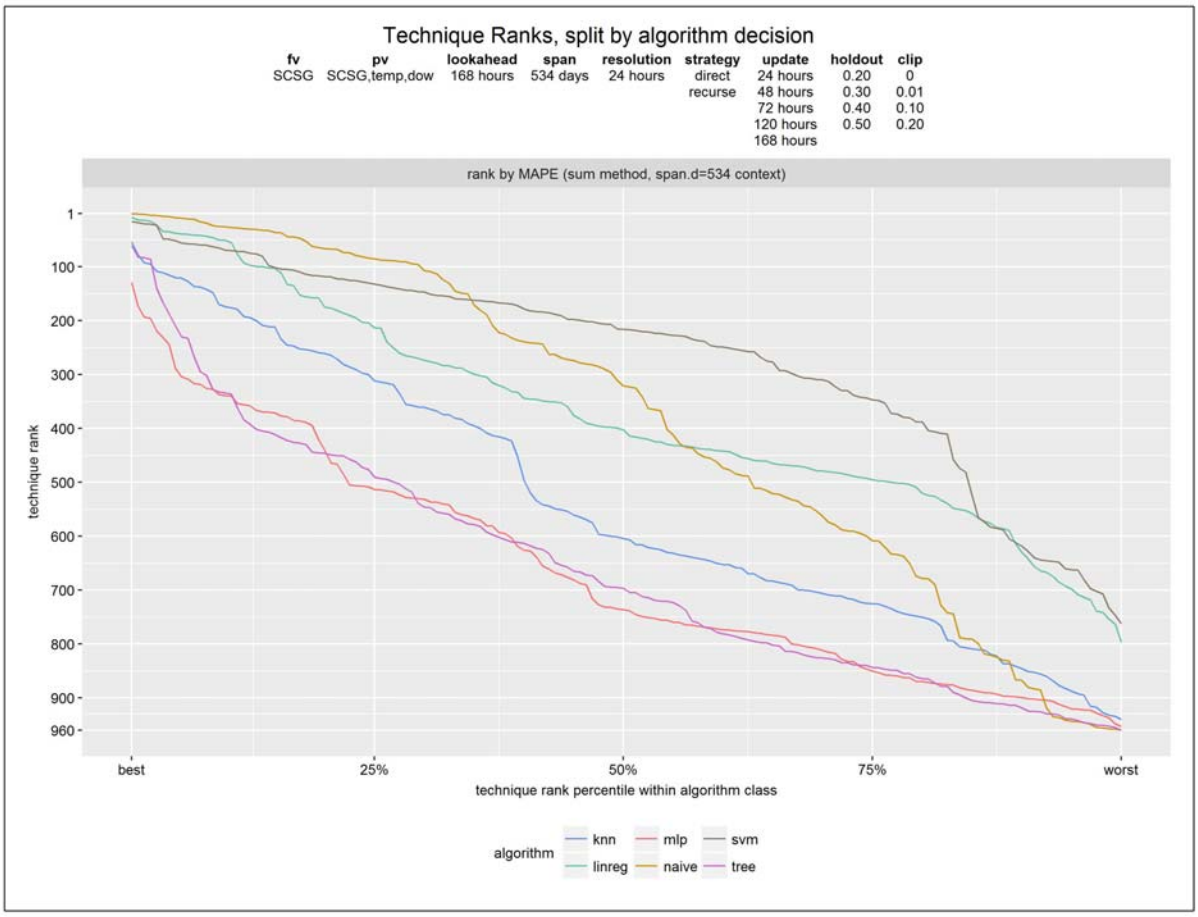


Figure 6-27: Technique rank trend at long span, split by algorithm class, Australia, week-ahead forecasts. 960 techniques and forecasts, all assume a decision for 534-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

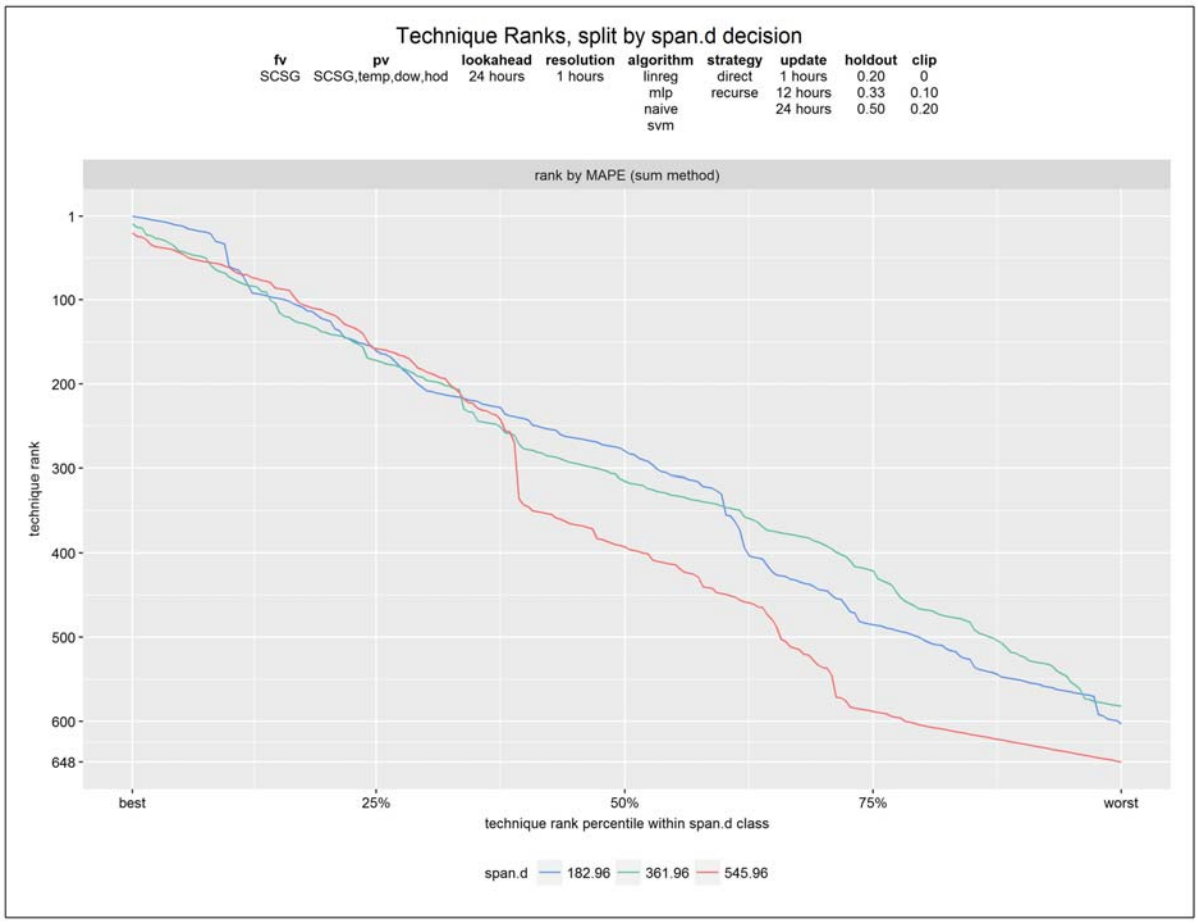


Figure 6-28: Technique rank trend, split by span, Australia, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the span decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the span decision option.

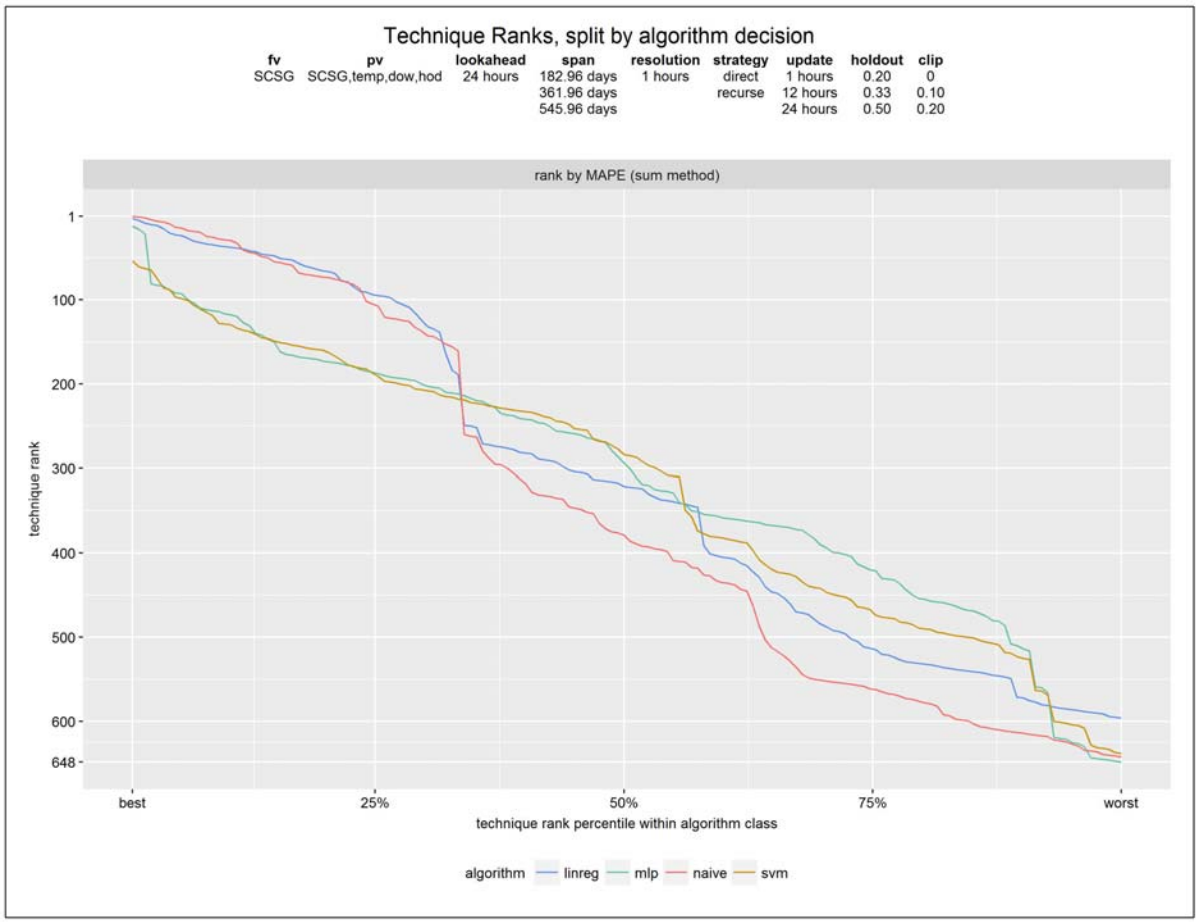


Figure 6-29: Technique rank trend, split by algorithm class, Australia, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

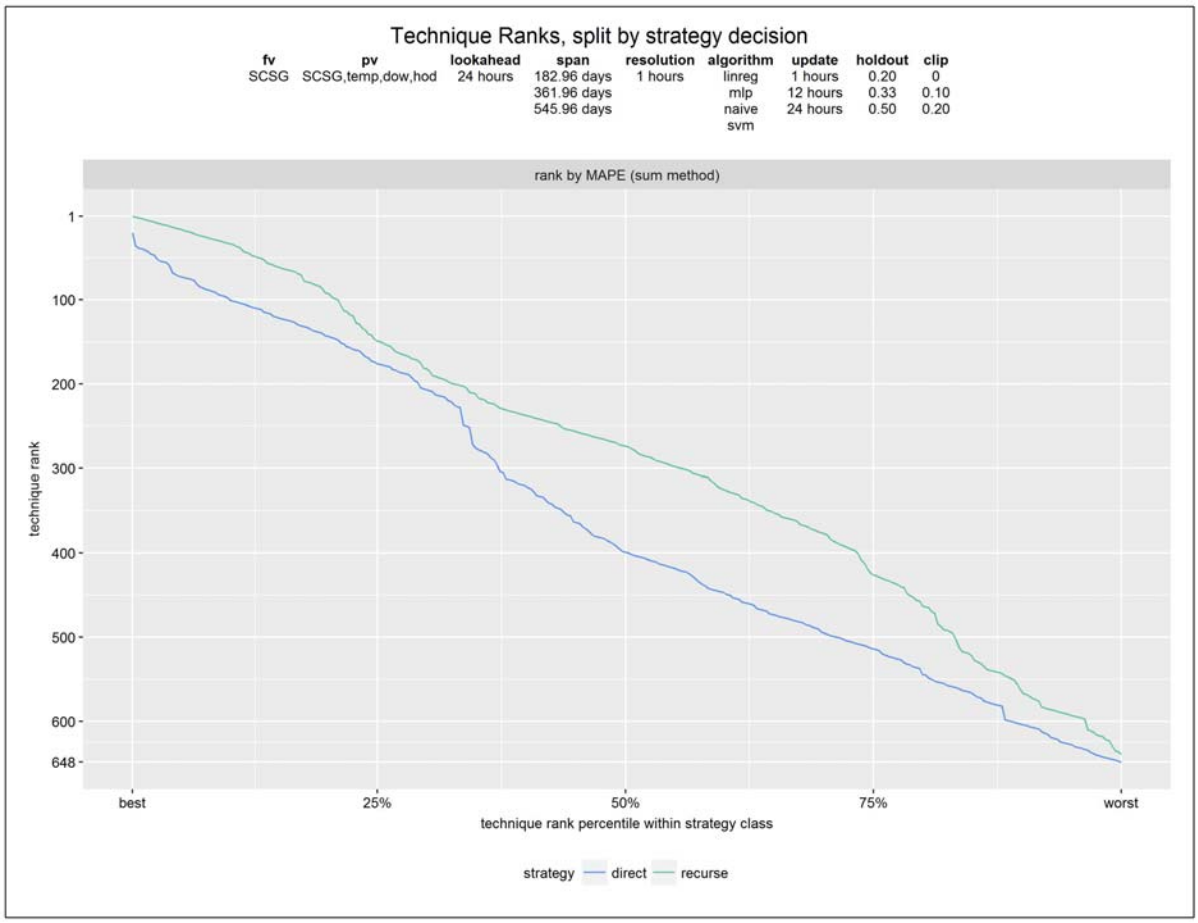


Figure 6-30: Technique rank trend, split by extension rule, Australia, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the extension rule decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the extension rule decision option.

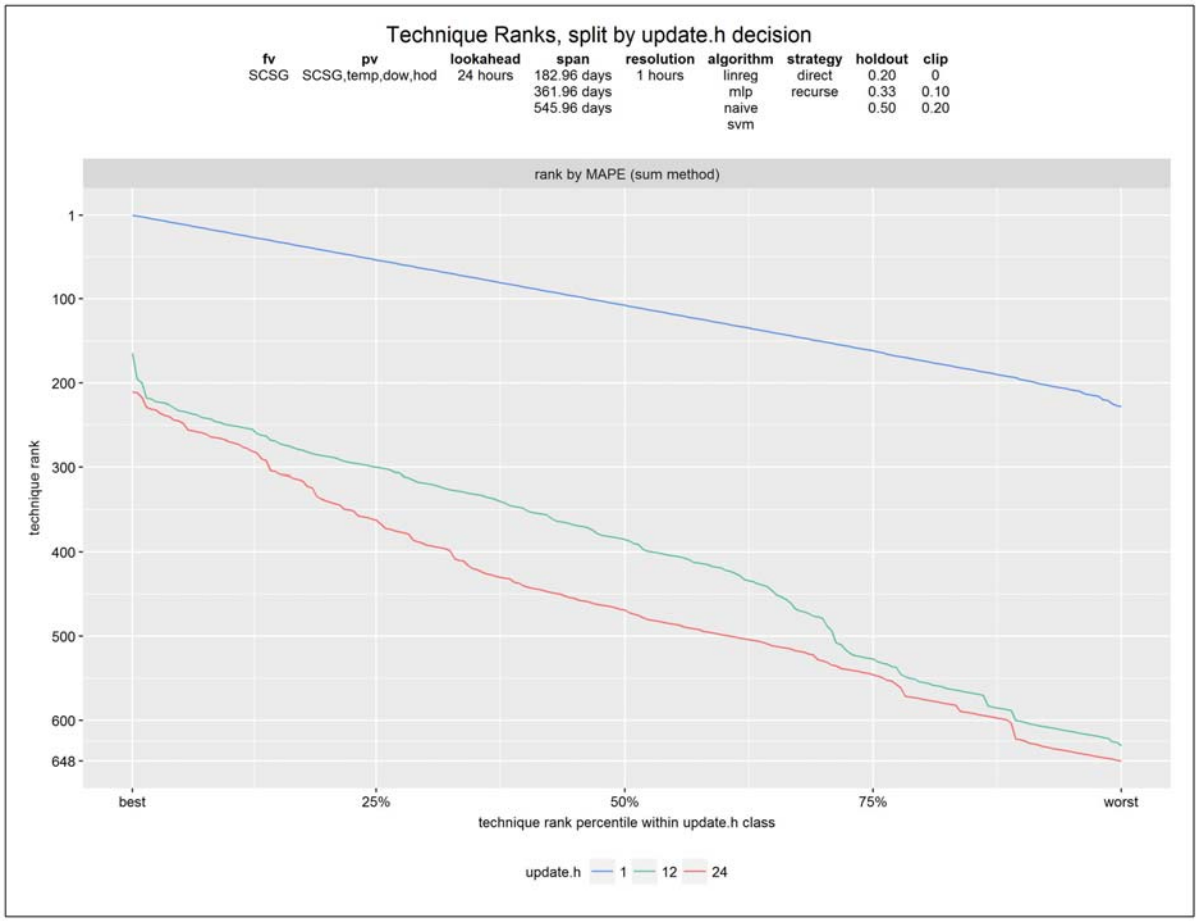


Figure 6-31: Technique rank trend, split by update cycle, Australia, day-ahead forecast. 648 techniques and forecasts. Each curve represents a family of techniques with the update cycle decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the update cycle decision option.

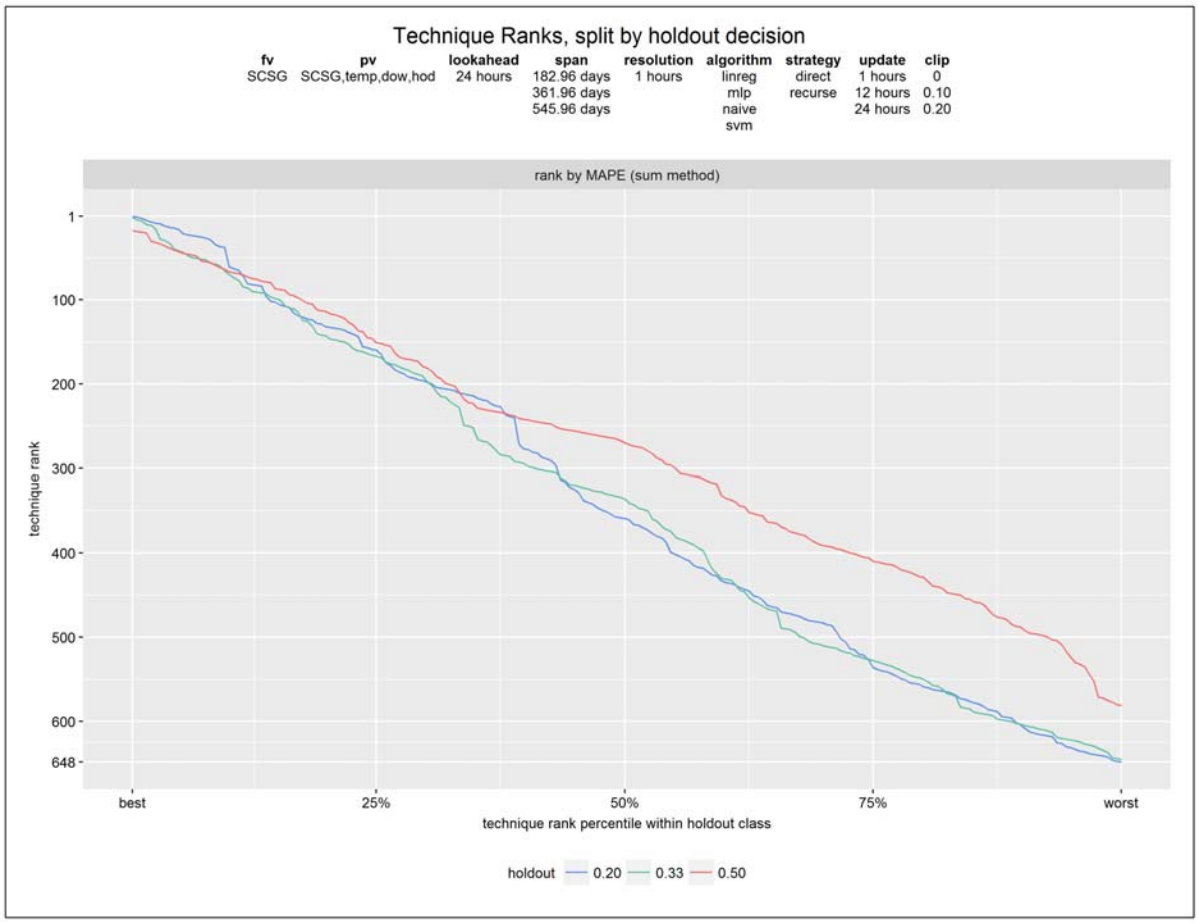


Figure 6-32: Technique rank trend, split by holdout, Australia, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the holdout decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the holdout decision option.

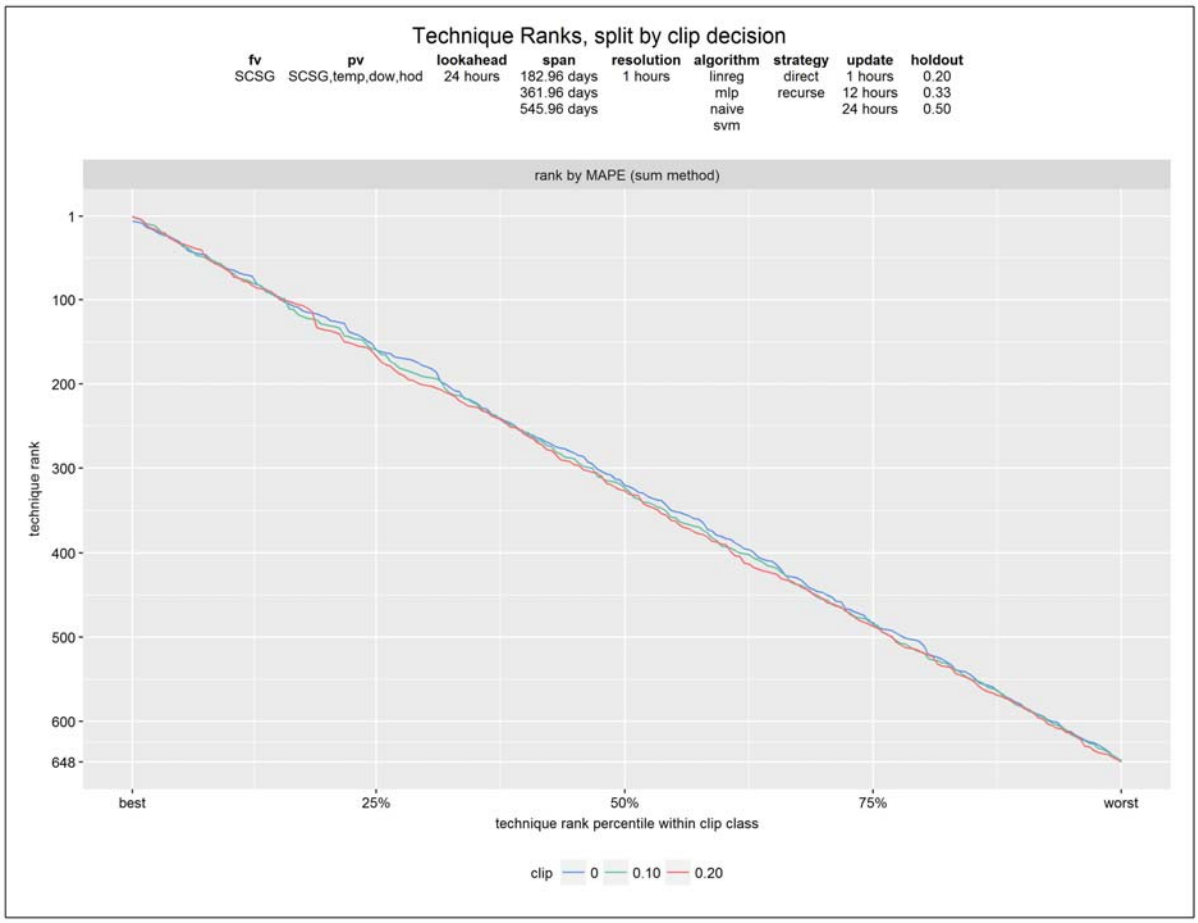


Figure 6-33: Technique rank trend, split by clip, Australia, day-ahead forecasts. 648 techniques and forecasts. Each curve represents a family of techniques with the clip decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the clip decision option.

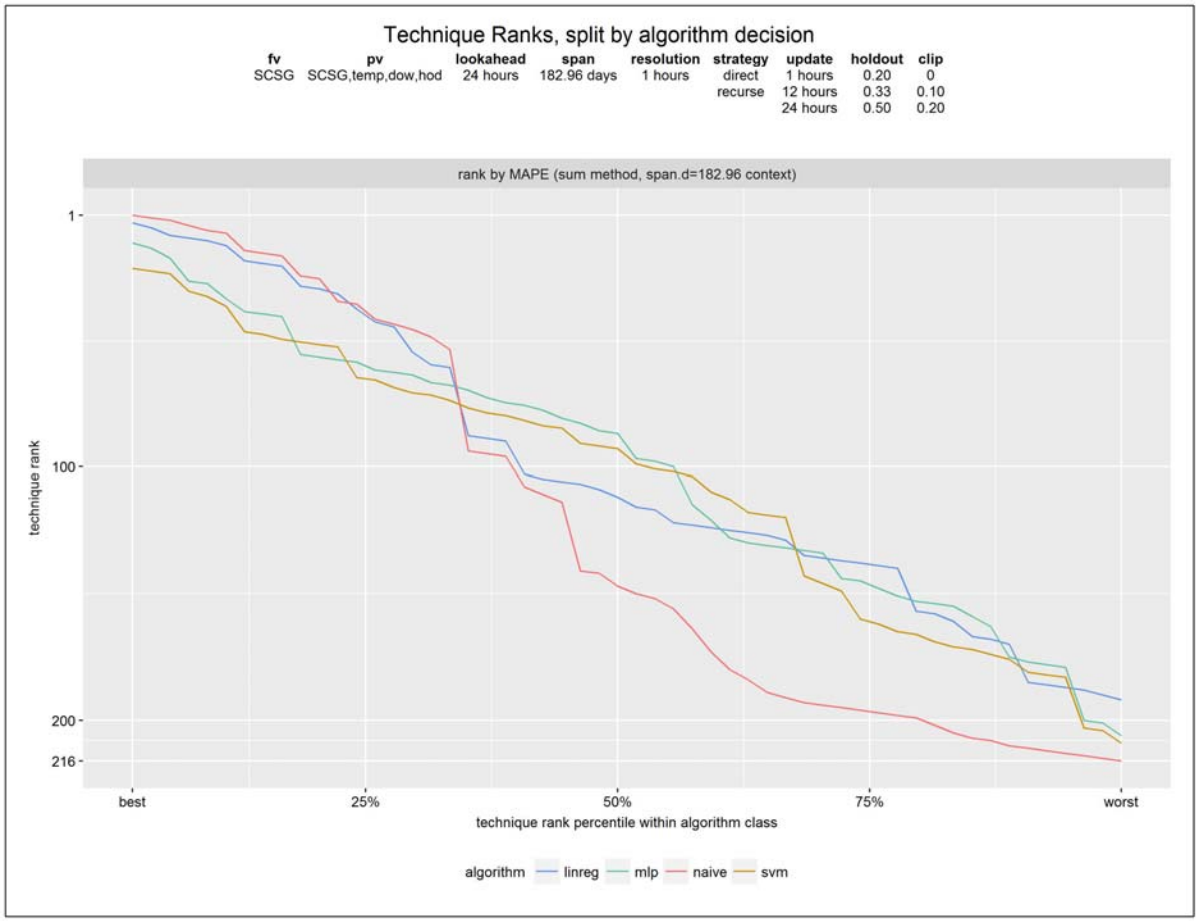


Figure 6-34: Technique rank trend at short span, split by algorithm class, Australia, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 182-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

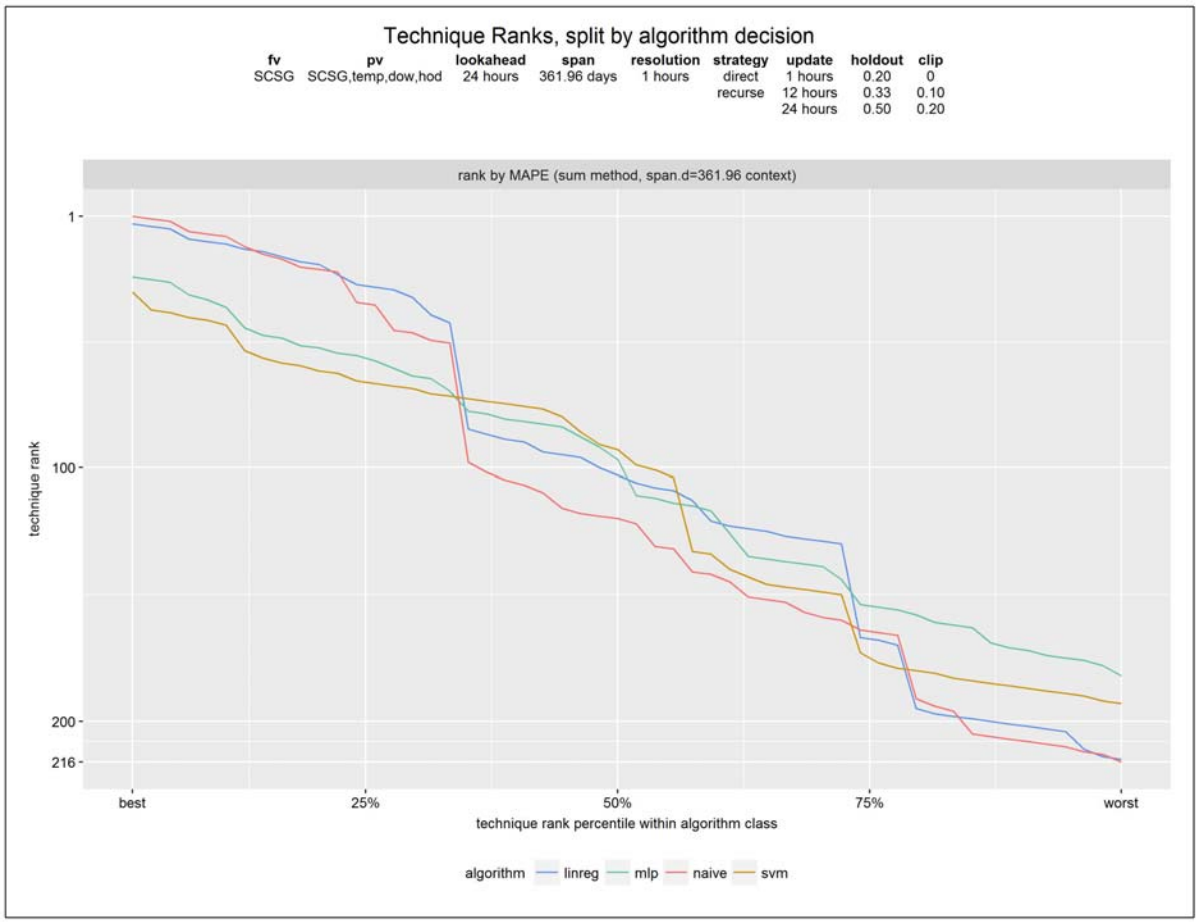


Figure 6-35: Technique rank trend at medium span, split by algorithm class, Australia, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 363-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.

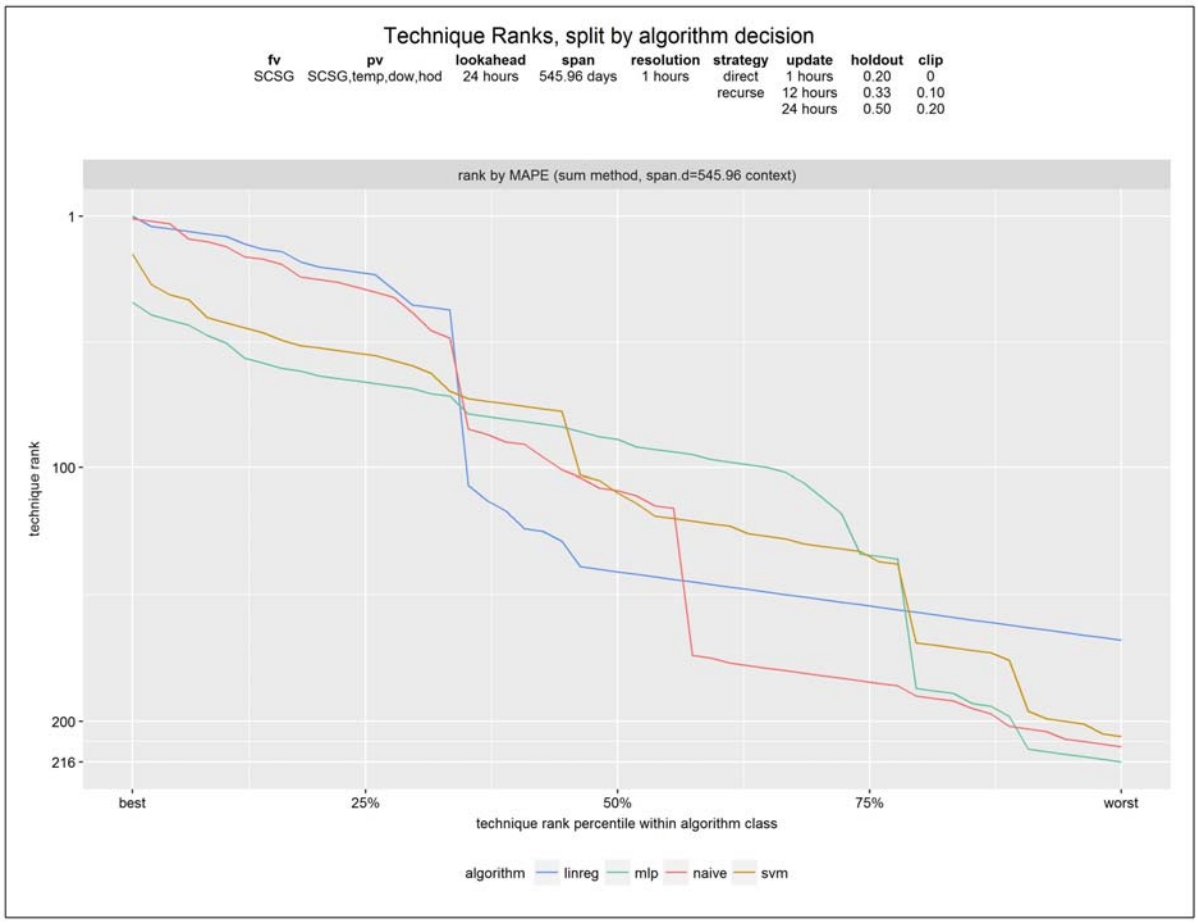
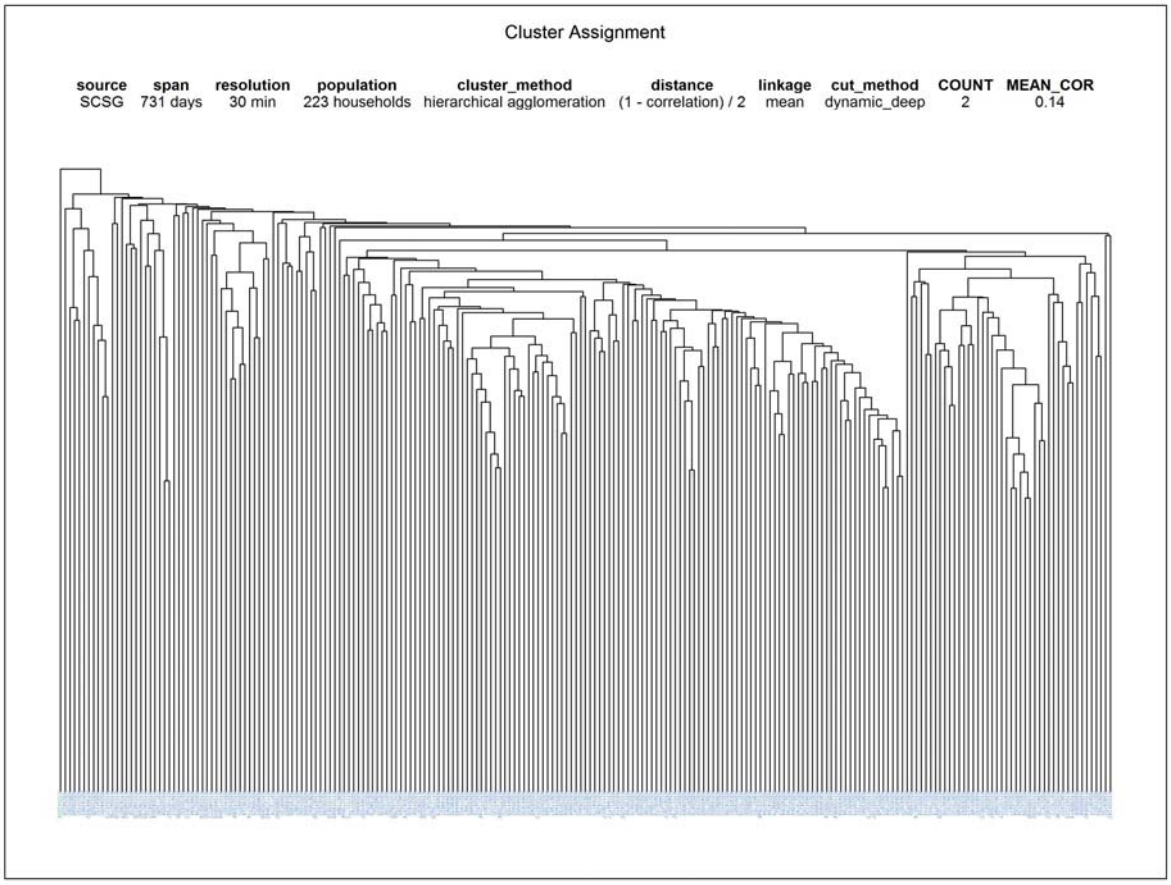
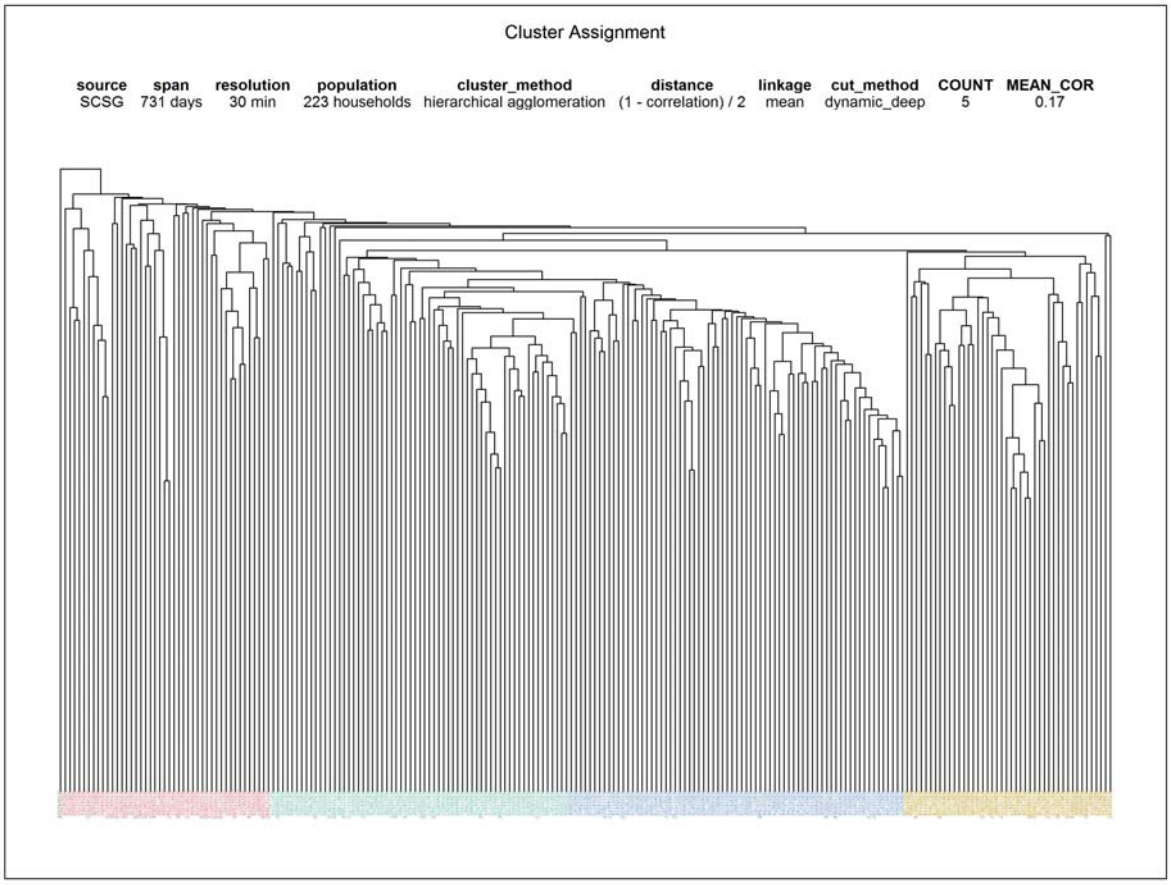


Figure 6-36: Technique rank trend at long span, split by algorithm class, Australia, day-ahead forecasts. 216 techniques and forecasts, all assume a decision for 534-day span. Each curve represents a family of techniques with the algorithm class decision option in common, each point on a curve represents a technique. Points are arranged along the y-axis in order of the technique rank within the population of techniques. Points are arranged along the x-axis in order of the technique percentile rank within the family of techniques represented by that curve. Metric is MAPE. *Colors* are the algorithm class decision option.



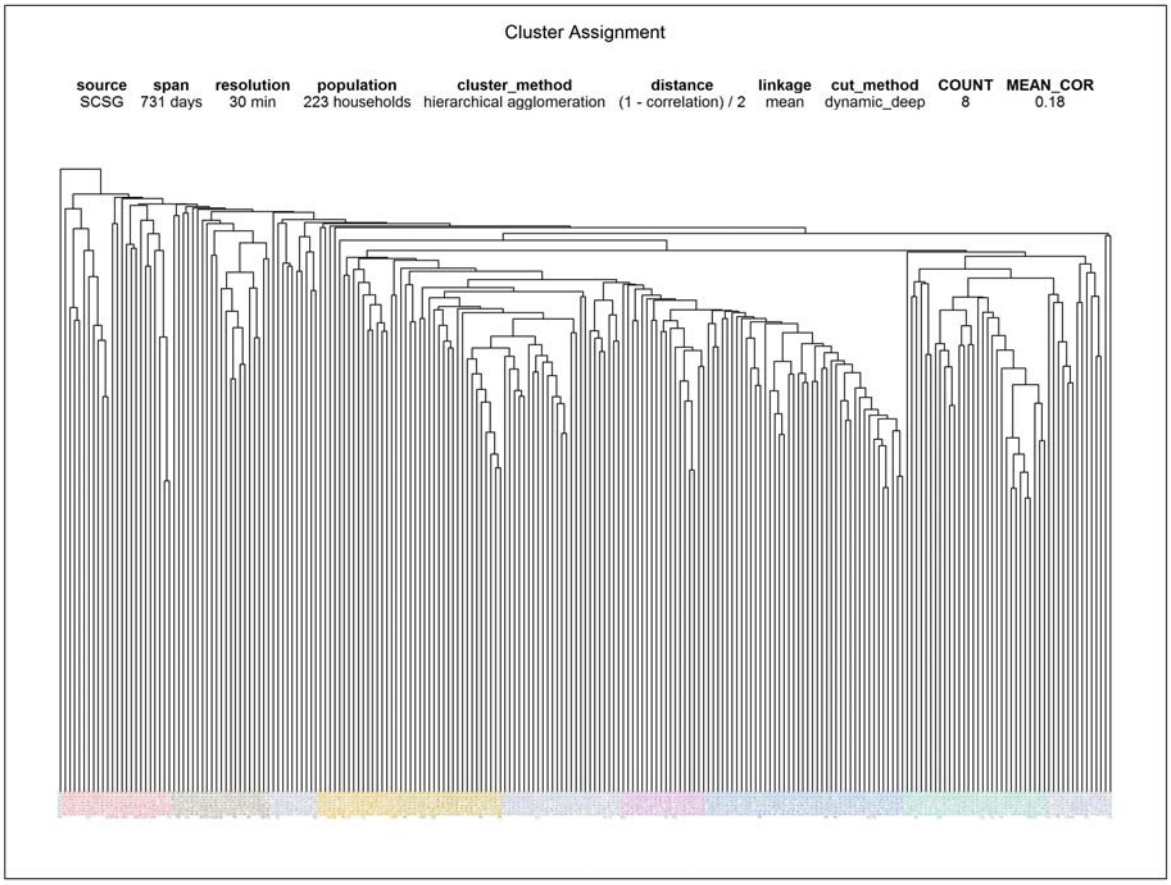

```
2 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  223 (distance mean 0.434, sd 0.046, max 0.649, min 0.232  :::  cor mean 0.131, sd 0.093, max 0.535, min -0.298)
2 clusters of sizes ...
  222 (distance mean 0.434, sd 0.046, max 0.649, min 0.232  :::  cor mean 0.133, sd 0.092, max 0.535, min -0.298)
    1 (distance mean 0.000, sd 0.000, max 0.000, min 0.000  :::  cor mean 1.000, sd 0.000, max 1.000, min 1.000)
                                                            weighted mean 0.136
```

Figure 6-37: Group 223 households as 2 clusters, Australia. *Colors* indicate to which clusters households are assigned.



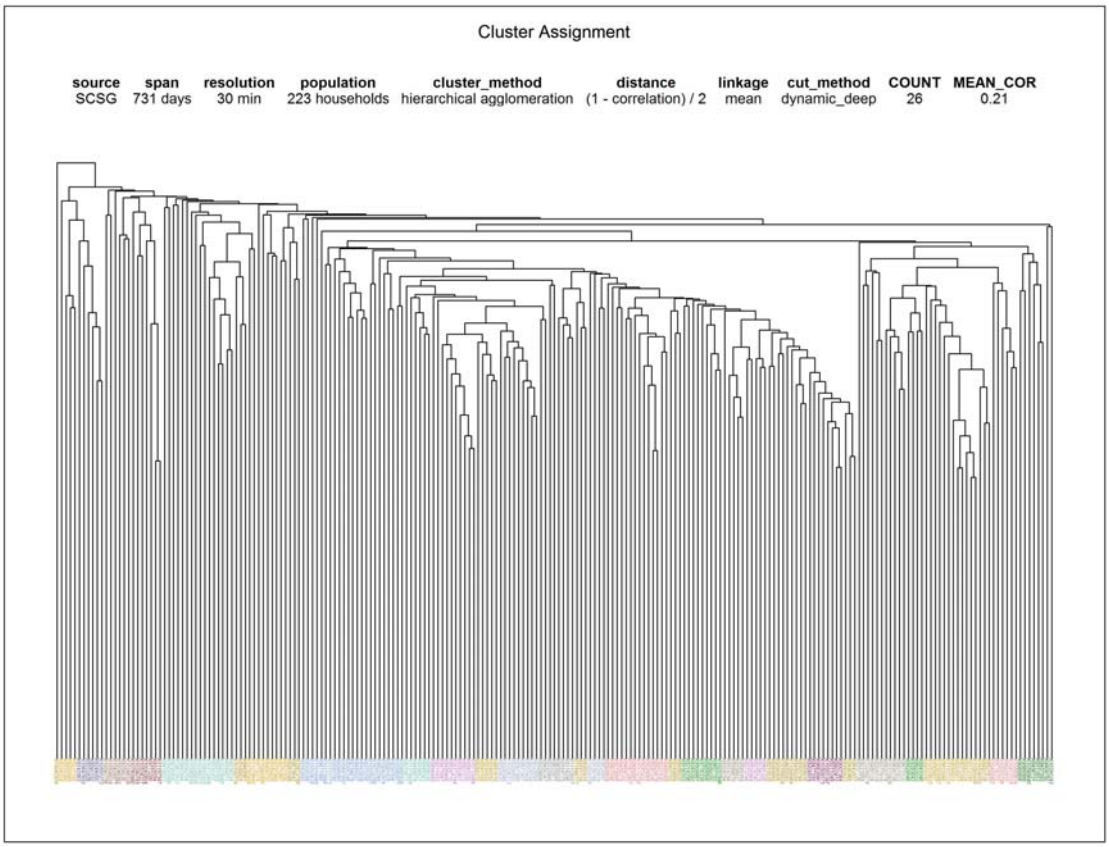

```
5 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  223 (distance mean 0.434, sd 0.046, max 0.649, min 0.232  :::  cor mean 0.131, sd 0.093, max 0.535, min -0.298)
5 clusters of sizes ...
   71 (distance mean 0.381, sd 0.033, max 0.495, min 0.241  :::  cor mean 0.238, sd 0.065, max 0.519, min 0.009)
   63 (distance mean 0.424, sd 0.044, max 0.561, min 0.256  :::  cor mean 0.152, sd 0.089, max 0.487, min -0.122)
   44 (distance mean 0.477, sd 0.041, max 0.649, min 0.246  :::  cor mean 0.046, sd 0.082, max 0.508, min -0.298)
   44 (distance mean 0.405, sd 0.043, max 0.497, min 0.232  :::  cor mean 0.191, sd 0.086, max 0.535, min 0.005)
    1 (distance mean 0.000, sd 0.000, max 0.000, min 0.000  :::  cor mean 1.000, sd 0.000, max 1.000, min 1.000)
                                                            weighted mean 0.170
```

Figure 6-38: Group 223 households as 4 clusters, Australia. *Colors* indicate to which clusters households are assigned.




```
8 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  223 (distance mean 0.434, sd 0.046, max 0.649, min 0.232  :::  cor mean 0.131, sd 0.093, max 0.535, min -0.298)
8 clusters of sizes ...
   42 (distance mean 0.361, sd 0.033, max 0.450, min 0.241  :::  cor mean 0.278, sd 0.066, max 0.519, min 0.100)
   30 (distance mean 0.386, sd 0.046, max 0.496, min 0.232  :::  cor mean 0.228, sd 0.091, max 0.535, min 0.007)
   23 (distance mean 0.483, sd 0.056, max 0.649, min 0.246  :::  cor mean 0.034, sd 0.112, max 0.508, min -0.298)
   22 (distance mean 0.391, sd 0.043, max 0.475, min 0.256  :::  cor mean 0.218, sd 0.086, max 0.487, min 0.051)
   21 (distance mean 0.451, sd 0.038, max 0.525, min 0.326  :::  cor mean 0.098, sd 0.076, max 0.347, min -0.051)
   18 (distance mean 0.388, sd 0.031, max 0.443, min 0.254  :::  cor mean 0.223, sd 0.062, max 0.491, min 0.114)
   17 (distance mean 0.429, sd 0.027, max 0.489, min 0.364  :::  cor mean 0.141, sd 0.054, max 0.272, min 0.022)
   50 (distance mean 0.429, sd 0.042, max 0.537, min 0.283  :::  cor mean 0.142, sd 0.084, max 0.434, min -0.074)
                                                          weighted mean 0.178
```

Figure 6-39: Group 223 households as 8 clusters, Australia. *Colors* indicate to which clusters households are assigned.




```
26 CLUSTERS BASED ON correlation , dynamic_deep tree cut:
distance measure is (1 - correlation) / 2
population of size  ...
  223 (distance mean 0.434, sd 0.046, max 0.649, min 0.232  :::  cor mean 0.131, sd 0.093, max 0.535, min -0.298)
26 clusters of sizes ...
   17 (distance mean 0.429, sd 0.027, max 0.489, min 0.364  :::  cor mean 0.141, sd 0.054, max 0.272, min 0.022)
   16 (distance mean 0.458, sd 0.039, max 0.525, min 0.326  :::  cor mean 0.084, sd 0.078, max 0.347, min -0.051)
   14 (distance mean 0.376, sd 0.032, max 0.436, min 0.254  :::  cor mean 0.248, sd 0.063, max 0.491, min 0.127)
   12 (distance mean 0.341, sd 0.053, max 0.431, min 0.232  :::  cor mean 0.318, sd 0.106, max 0.535, min 0.138)
   11 (distance mean 0.409, sd 0.039, max 0.481, min 0.305  :::  cor mean 0.183, sd 0.078, max 0.389, min 0.039)
   10 (distance mean 0.339, sd 0.040, max 0.405, min 0.256  :::  cor mean 0.323, sd 0.081, max 0.487, min 0.189)
    9 (distance mean 0.468, sd 0.029, max 0.530, min 0.412  :::  cor mean 0.065, sd 0.057, max 0.177, min -0.060)
    9 (distance mean 0.337, sd 0.022, max 0.379, min 0.283  :::  cor mean 0.325, sd 0.044, max 0.434, min 0.243)
    9 (distance mean 0.385, sd 0.025, max 0.419, min 0.321  :::  cor mean 0.229, sd 0.049, max 0.358, min 0.163)
    9 (distance mean 0.356, sd 0.026, max 0.415, min 0.293  :::  cor mean 0.288, sd 0.052, max 0.413, min 0.170)
    8 (distance mean 0.405, sd 0.032, max 0.455, min 0.348  :::  cor mean 0.190, sd 0.064, max 0.304, min 0.090)
    8 (distance mean 0.303, sd 0.025, max 0.349, min 0.241  :::  cor mean 0.393, sd 0.049, max 0.519, min 0.302)
    8 (distance mean 0.420, sd 0.028, max 0.476, min 0.344  :::  cor mean 0.160, sd 0.056, max 0.312, min 0.049)
    7 (distance mean 0.424, sd 0.050, max 0.470, min 0.246  :::  cor mean 0.152, sd 0.099, max 0.508, min 0.060)
    6 (distance mean 0.393, sd 0.044, max 0.458, min 0.312  :::  cor mean 0.213, sd 0.088, max 0.376, min 0.084)
    6 (distance mean 0.473, sd 0.024, max 0.503, min 0.430  :::  cor mean 0.054, sd 0.048, max 0.140, min -0.006)
    6 (distance mean 0.431, sd 0.031, max 0.468, min 0.372  :::  cor mean 0.138, sd 0.062, max 0.257, min 0.064)
    6 (distance mean 0.385, sd 0.025, max 0.428, min 0.351  :::  cor mean 0.230, sd 0.050, max 0.298, min 0.144)
    6 (distance mean 0.372, sd 0.028, max 0.421, min 0.323  :::  cor mean 0.256, sd 0.056, max 0.354, min 0.157)
    5 (distance mean 0.333, sd 0.013, max 0.350, min 0.312  :::  cor mean 0.335, sd 0.025, max 0.375, min 0.301)
    5 (distance mean 0.333, sd 0.038, max 0.385, min 0.282  :::  cor mean 0.333, sd 0.077, max 0.436, min 0.230)
    5 (distance mean 0.342, sd 0.013, max 0.362, min 0.324  :::  cor mean 0.316, sd 0.026, max 0.353, min 0.276)
    4 (distance mean 0.423, sd 0.047, max 0.484, min 0.373  :::  cor mean 0.154, sd 0.093, max 0.254, min 0.031)
    4 (distance mean 0.416, sd 0.022, max 0.439, min 0.373  :::  cor mean 0.169, sd 0.043, max 0.255, min 0.121)
    4 (distance mean 0.363, sd 0.013, max 0.388, min 0.353  :::  cor mean 0.274, sd 0.026, max 0.294, min 0.224)
   19 (distance mean 0.414, sd 0.054, max 0.533, min 0.250  :::  cor mean 0.171, sd 0.108, max 0.501, min -0.067)
                                                          weighted mean 0.212
```

Figure 6-40: Group 223 households as 26 clusters, Australia. *Colors* indicate to which clusters households are assigned.

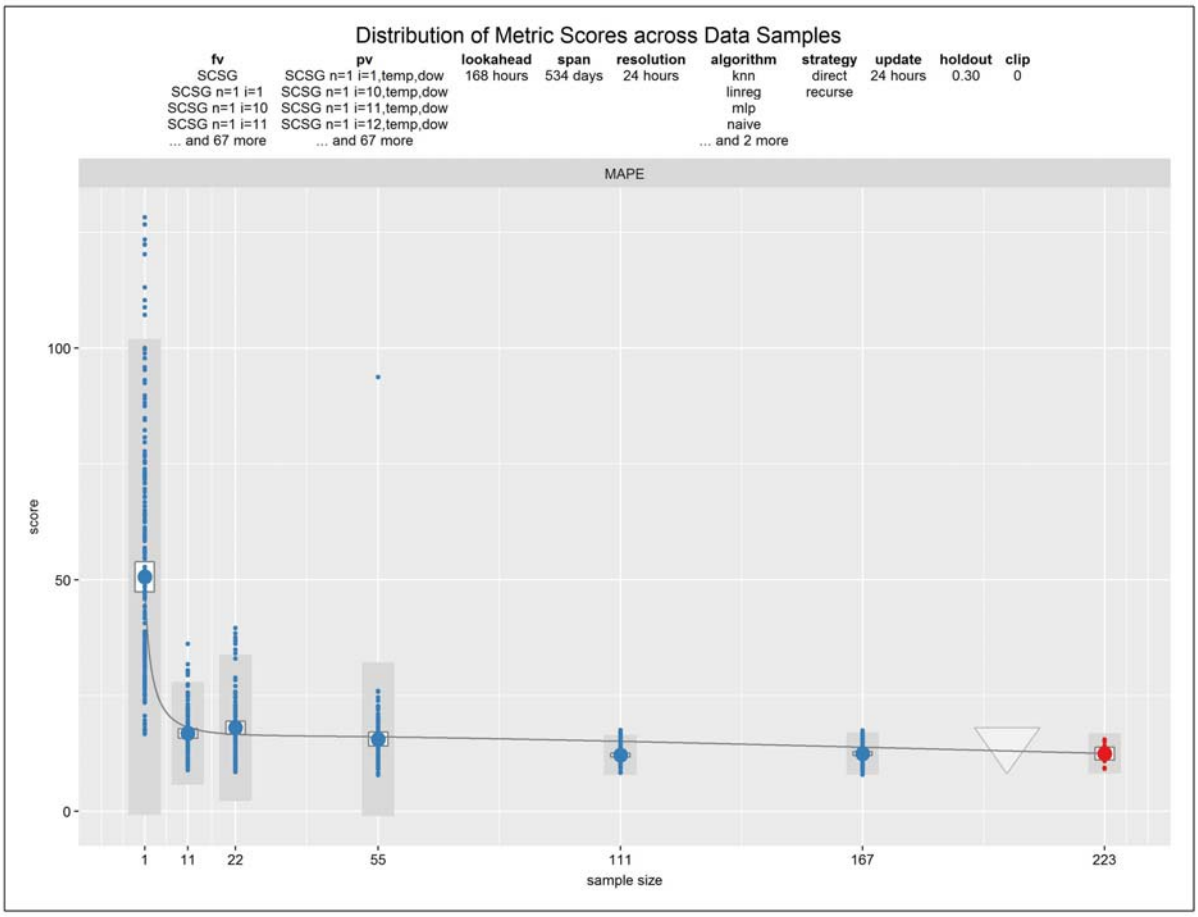


Figure 6-41: Effect of sampling – distribution of metric scores vs. sample size, Australia, week-ahead forecasts. 12 techniques. 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. *Blue* is metric score for a forecast at a specific sample size. *Large blue* is mean metric score of forecasts at a specific sample size. *Red* is metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

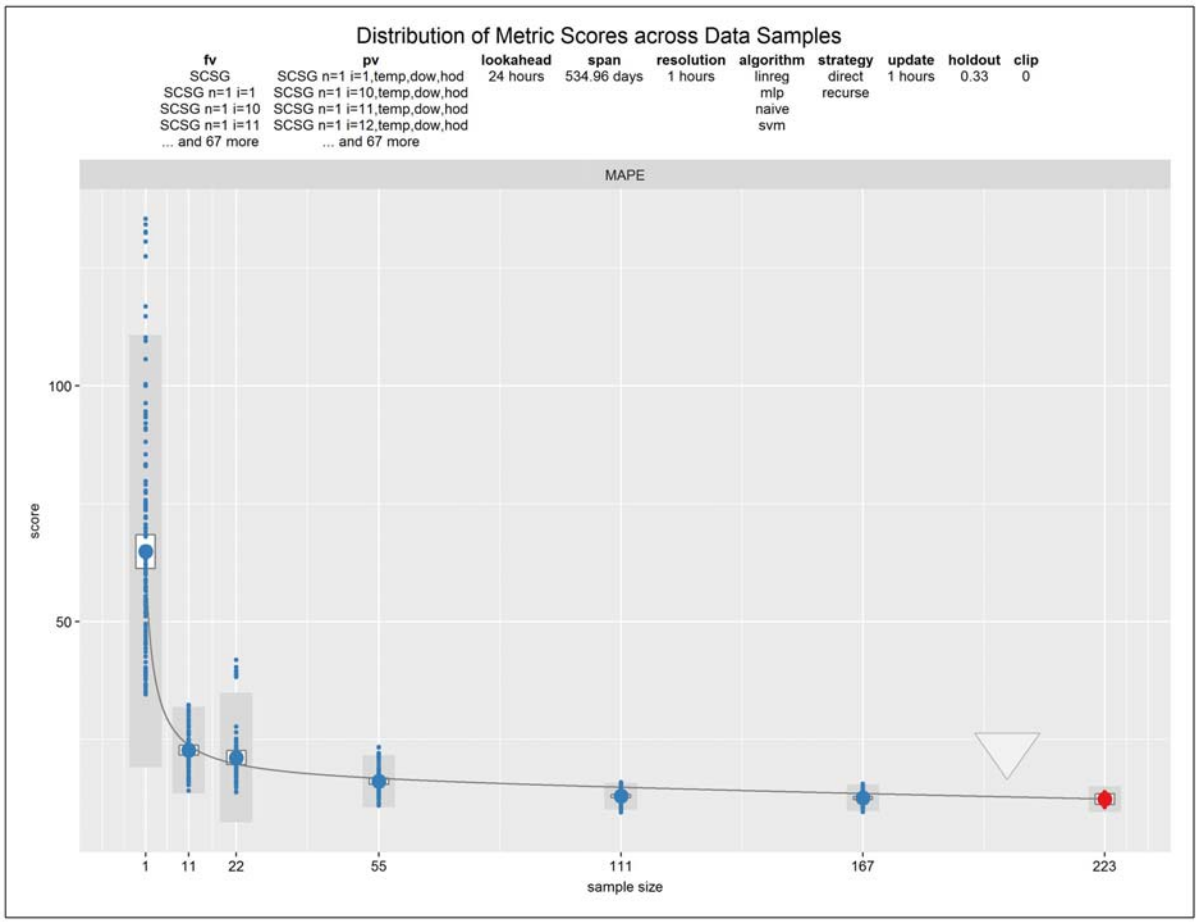


Figure 6-42: Effect of sampling – distribution of metric scores vs. sample size, Australia, day-ahead forecasts. 8 techniques. 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. *Blue* is metric score for a forecast at a specific sample size. *Large blue* is mean metric score of forecasts at a specific sample size. *Red* is metric score of a forecast at no sampling (full population). *Large red* is mean metric score of forecasts at no sampling (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific sample size. *White* is 95% confidence interval of mean metric score of forecasts at a specific sample size.

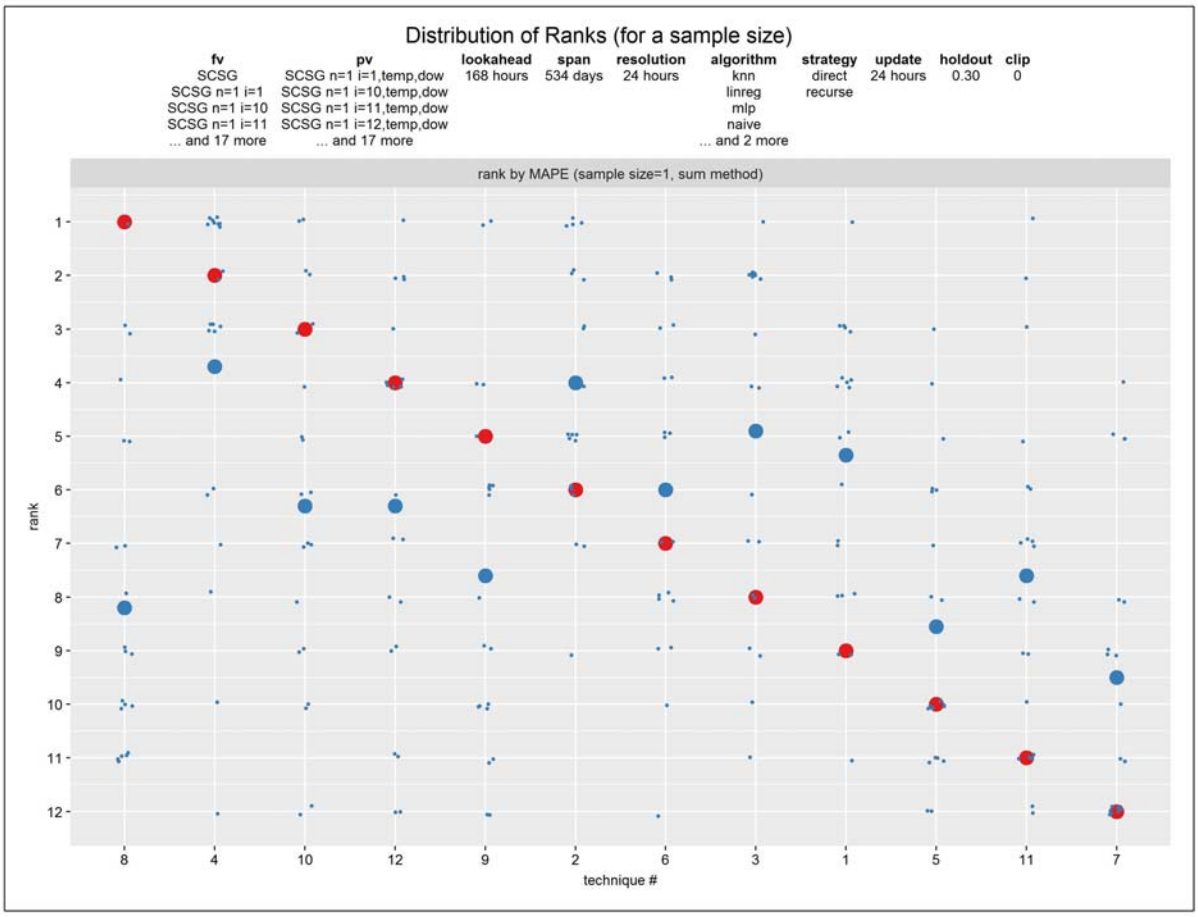


Figure 6-43: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 1, Australia, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 20 samples of size 1 out of 223. *Large blue* is mean technique rank for a specific technique applied to all 20 samples. *Large red* is technique rank for a specific technique applied to full population.

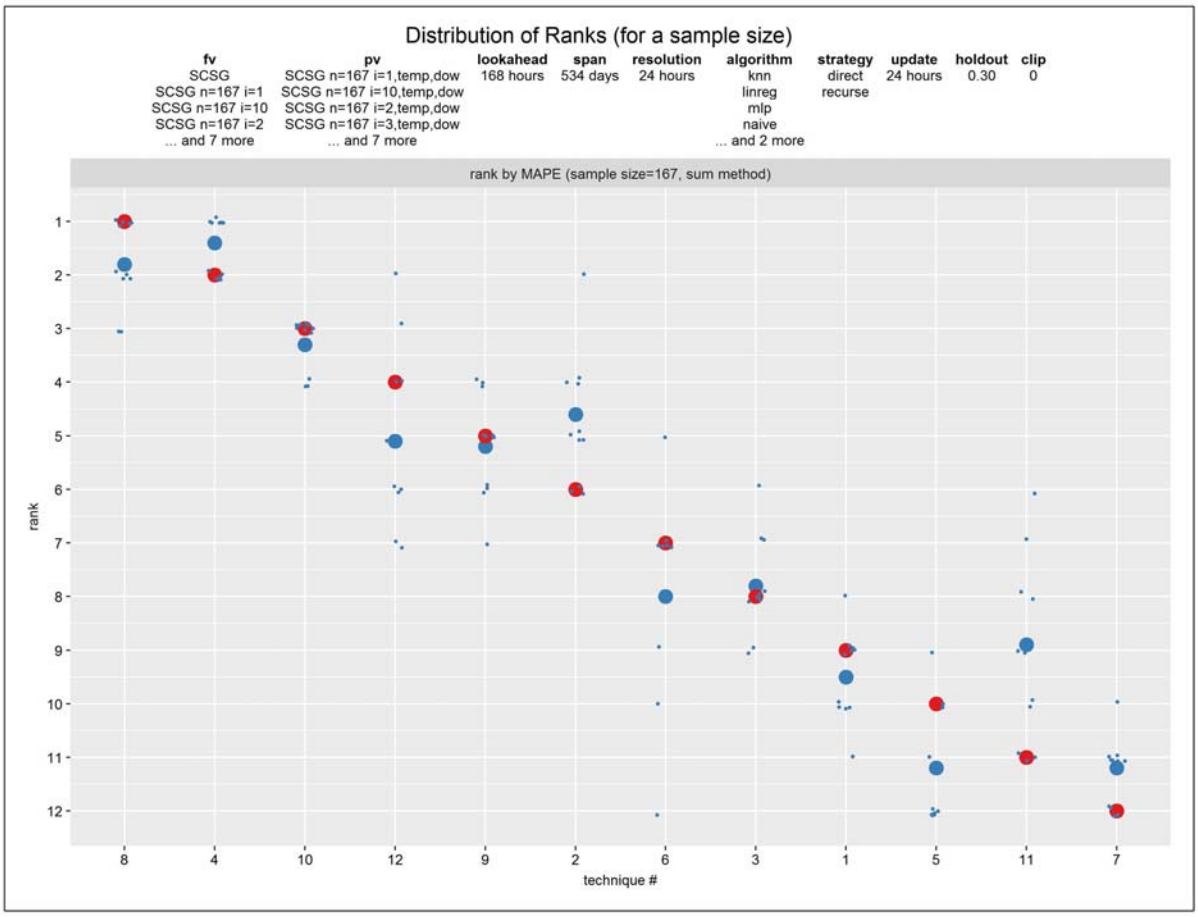


Figure 6-44: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 167, Australia, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 167 out of 223. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

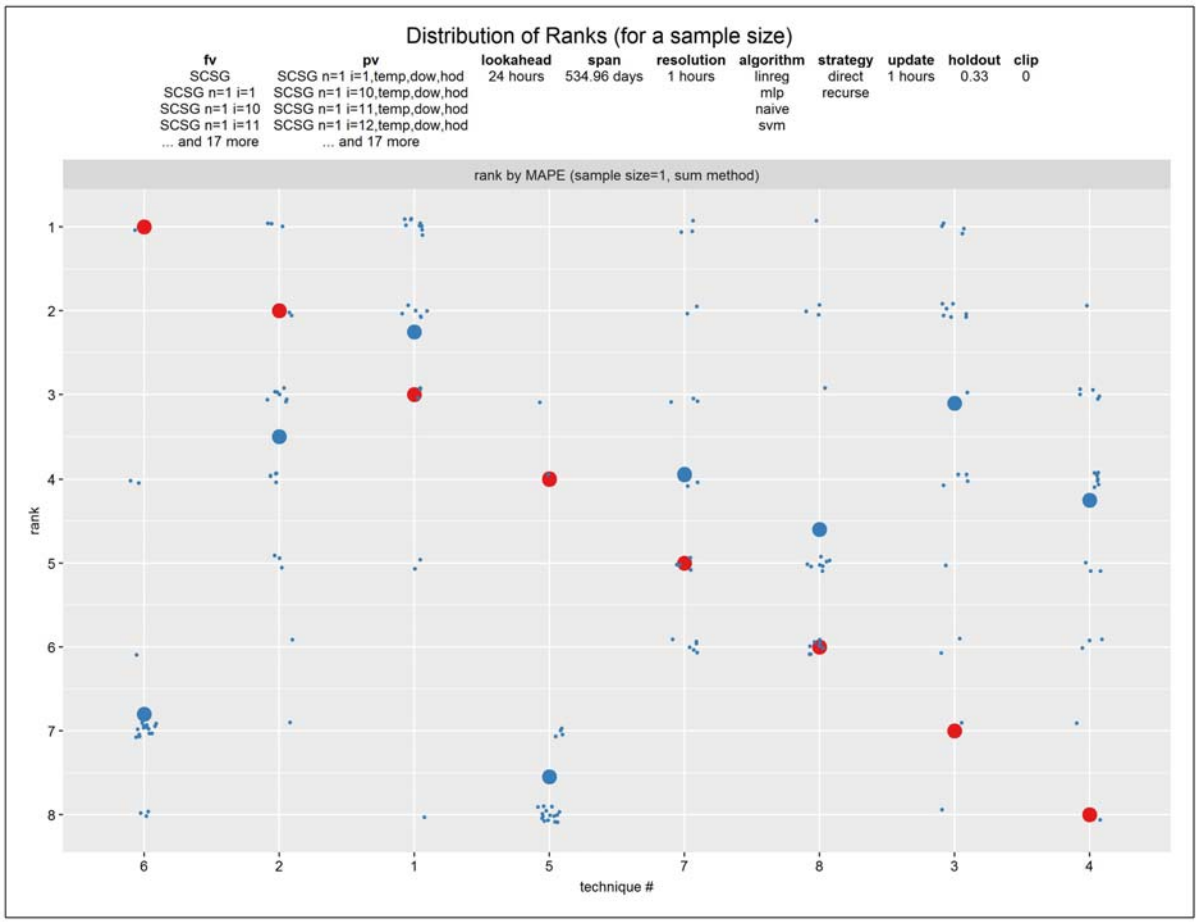


Figure 6-45: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 1, Australia, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 20 samples of size 1 out of 223. *Large blue* is mean technique rank for a specific technique applied to all 20 samples. *Large red* is technique rank for a specific technique applied to full population.

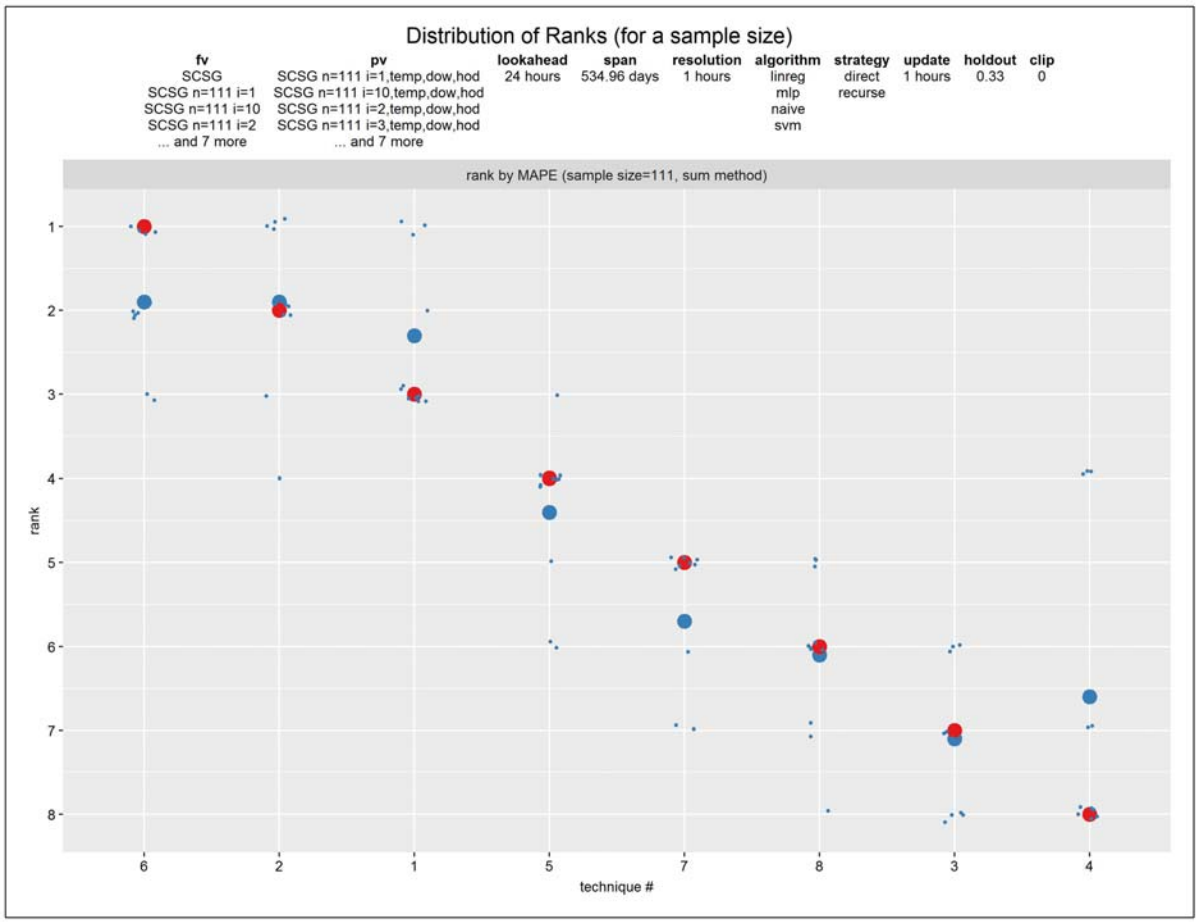


Figure 6-46: Effect of sampling – distribution of technique ranks across samples vs. technique at sample size 111, Australia, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied to 1 of 10 samples of size 111 out of 223. *Large blue* is mean technique rank for a specific technique applied to all 10 samples. *Large red* is technique rank for a specific technique applied to full population.

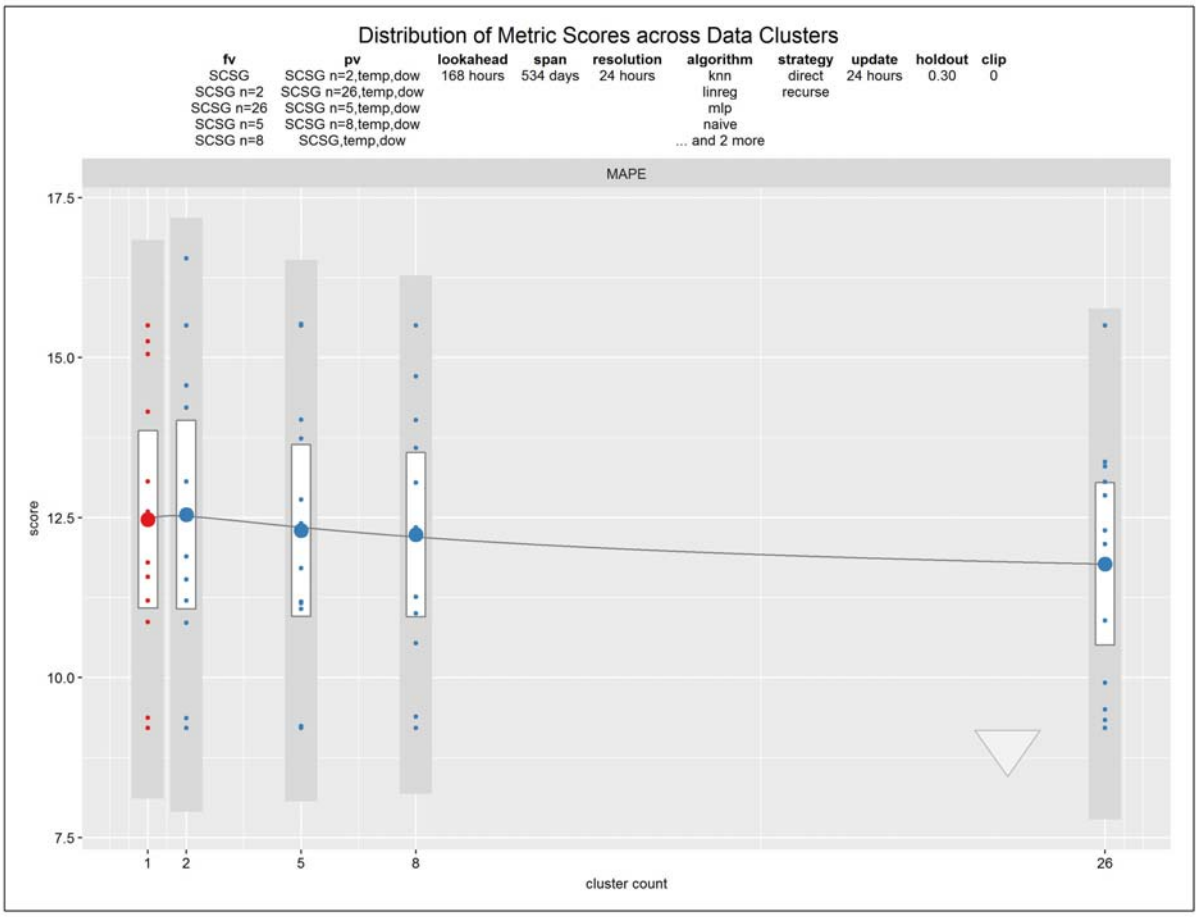


Figure 6-47: Effect of clustering – distribution of metric scores vs. number of clusters, Australia, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is metric score for a forecast at a specific number of clusters. *Large blue* is mean metric score of forecasts at a specific number of clusters. *Red* is metric score of a forecast at no clustering (full population). *Large red* is mean metric score of forecasts at no clustering (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific number of clusters. *White* is 95% confidence interval of mean metric score of forecasts at a specific number of clusters.

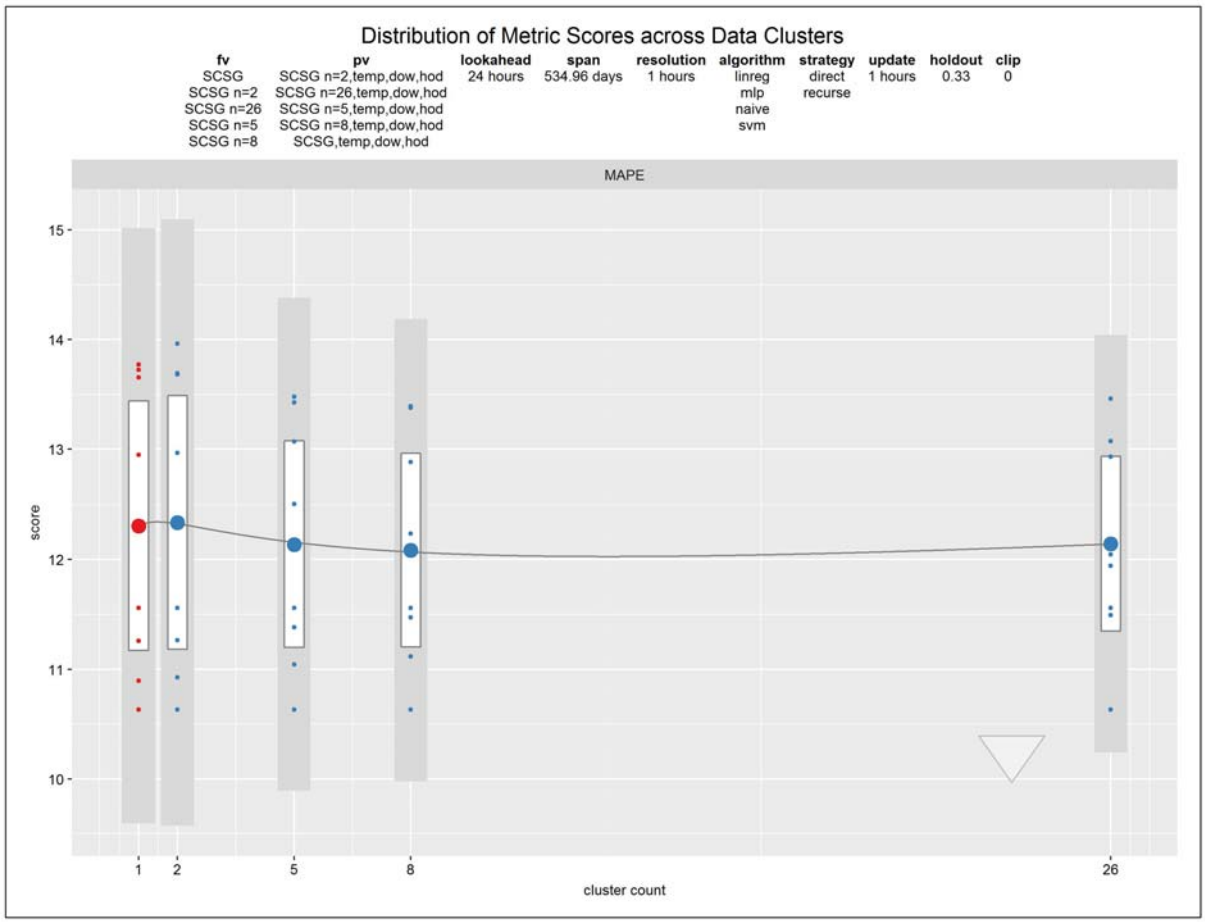


Figure 6-48: Effect of clustering – distribution of metric scores vs. number of clusters, Australia, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is metric score for a forecast at a specific number of clusters. *Large blue* is mean metric score of forecasts at a specific number of clusters. *Red* is metric score of a forecast at no clustering (full population). *Large red* is mean metric score of forecasts at no clustering (full population). *Gray* is 2 standard deviations from mean metric score of forecasts at a specific number of clusters. *White* is 95% confidence interval of mean metric score of forecasts at a specific number of clusters.

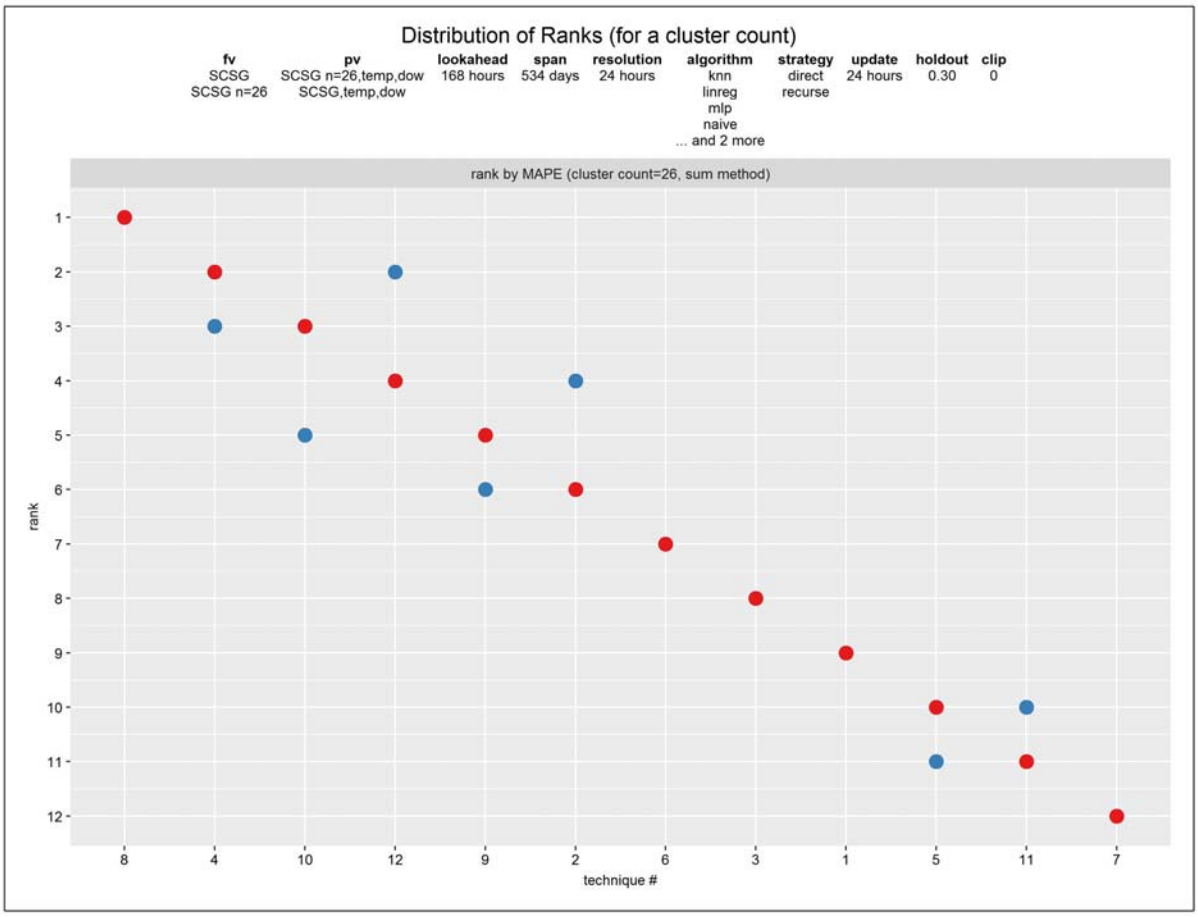


Figure 6-49: Effect of clustering – technique rank vs. technique at 26 clusters, Ireland, week-ahead forecasts. 12 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 26 clusters. *Red* is technique rank for a specific technique applied without clustering.

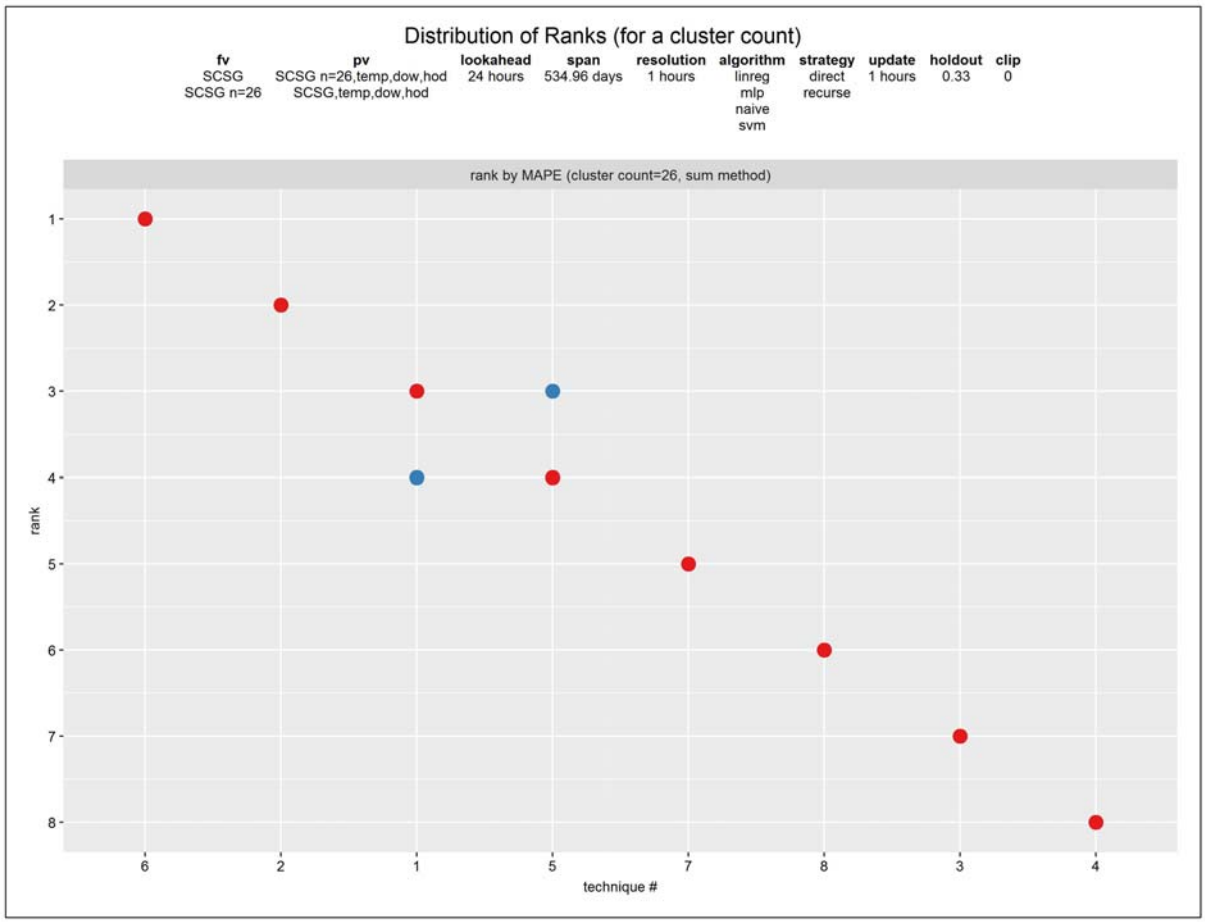


Figure 6-50: Effect of clustering – technique rank vs. technique at 26 clusters, Ireland, day-ahead forecasts. 8 techniques. Metric is MAPE. *Blue* is technique rank for a specific technique applied when grouping as 26 clusters. *Red* is technique rank for a specific technique applied without clustering.

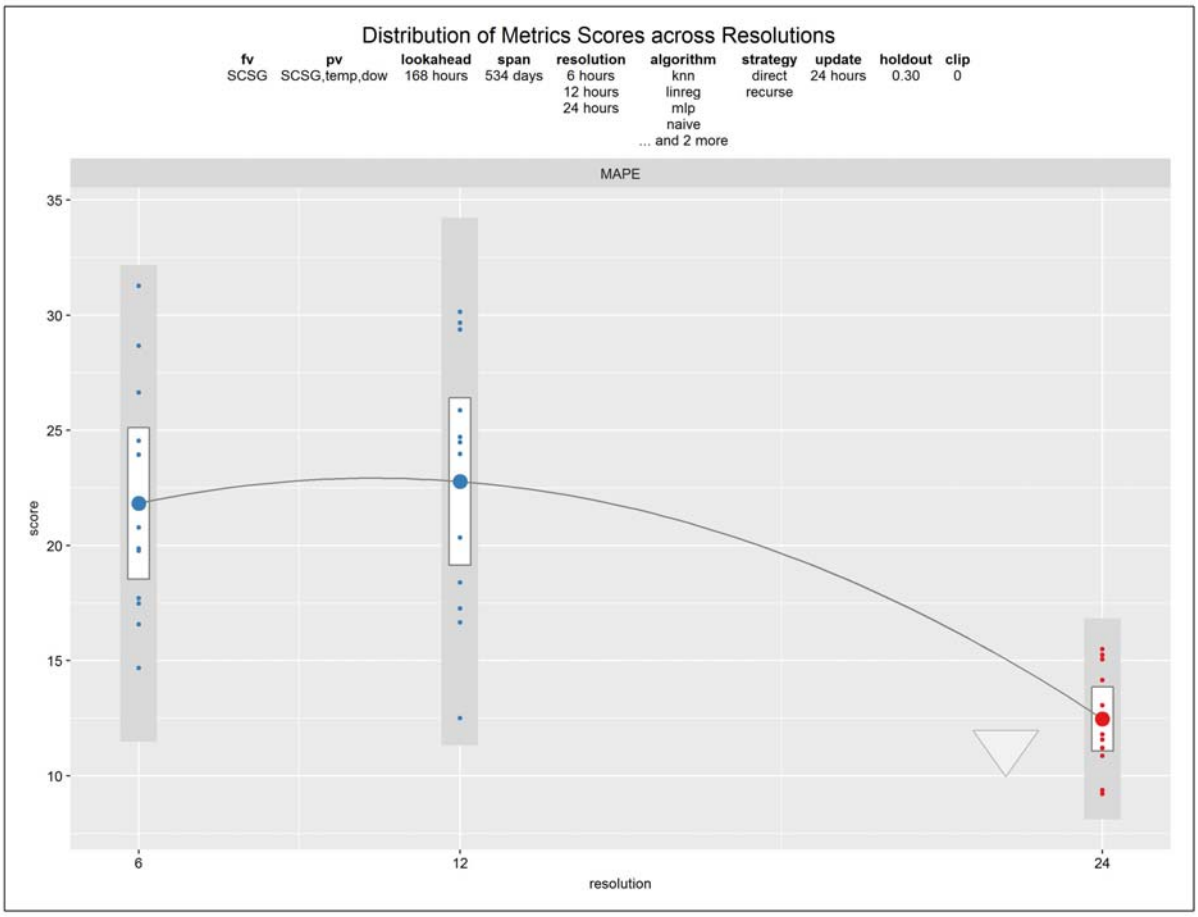


Figure 6-51: Effect of temporal magnification – distribution of metric scores vs. time step size, Australia, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. *Blue* is metric score for a forecast at a specific time step size. *Large blue* is mean metric score of forecasts at a specific time step size. *Red* is metric score of a forecast at baseline time step size. *Large red* is mean metric score of forecasts at baseline time step size. *Gray* is 2 standard deviations from mean metric score of forecasts at a specific time step size. *White* is 95% confidence interval of mean metric score of forecasts at a specific time step size.

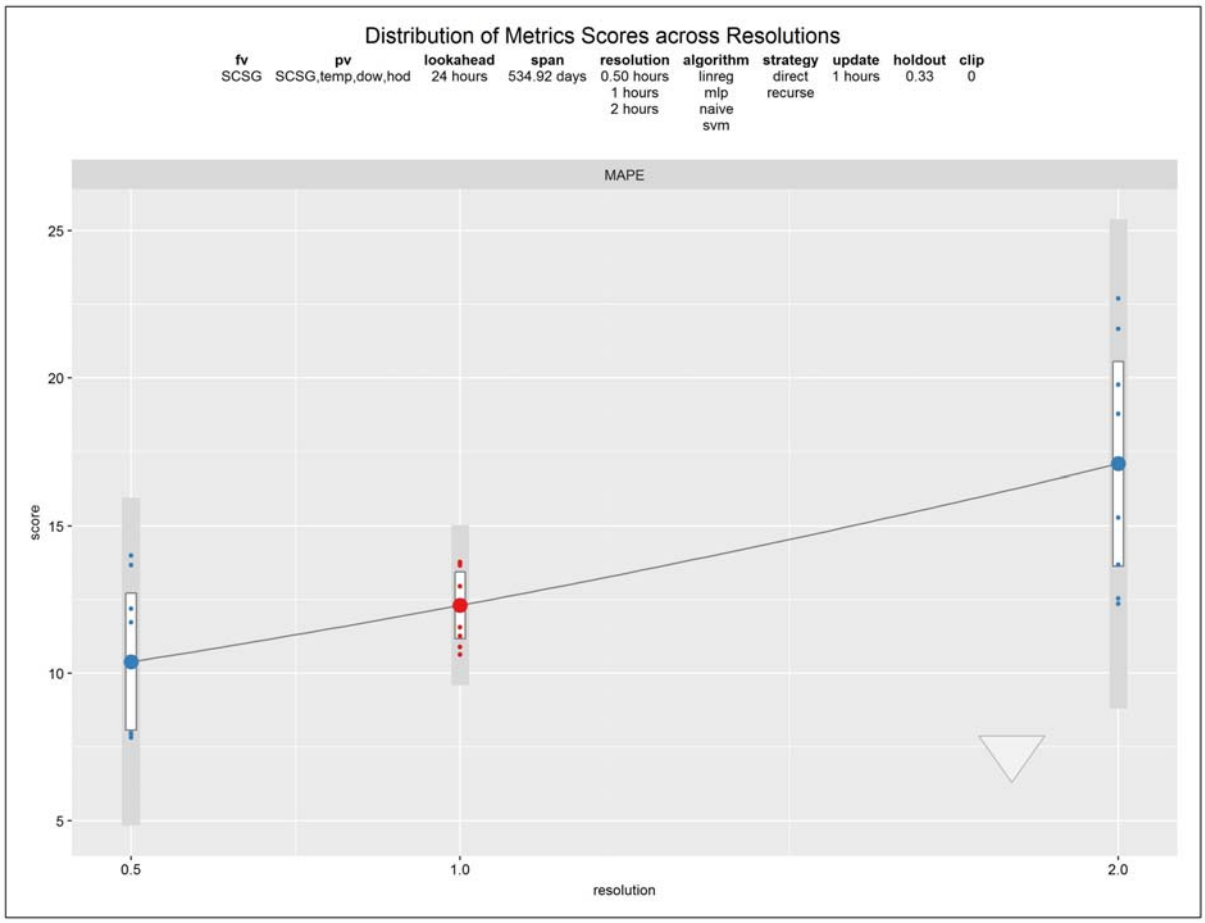


Figure 6-52: Effect of temporal magnification – distribution of metric scores vs. time step size, Australia, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. *Blue* is metric score for a forecast at a specific time step size. *Large blue* is mean metric score of forecasts at a specific time step size. *Red* is metric score of a forecast at baseline time step size. *Large red* is mean metric score of forecasts at baseline time step size. *Gray* is 2 standard deviations from mean metric score of forecasts at a specific time step size. *White* is 95% confidence interval of mean metric score of forecasts at a specific time step size.

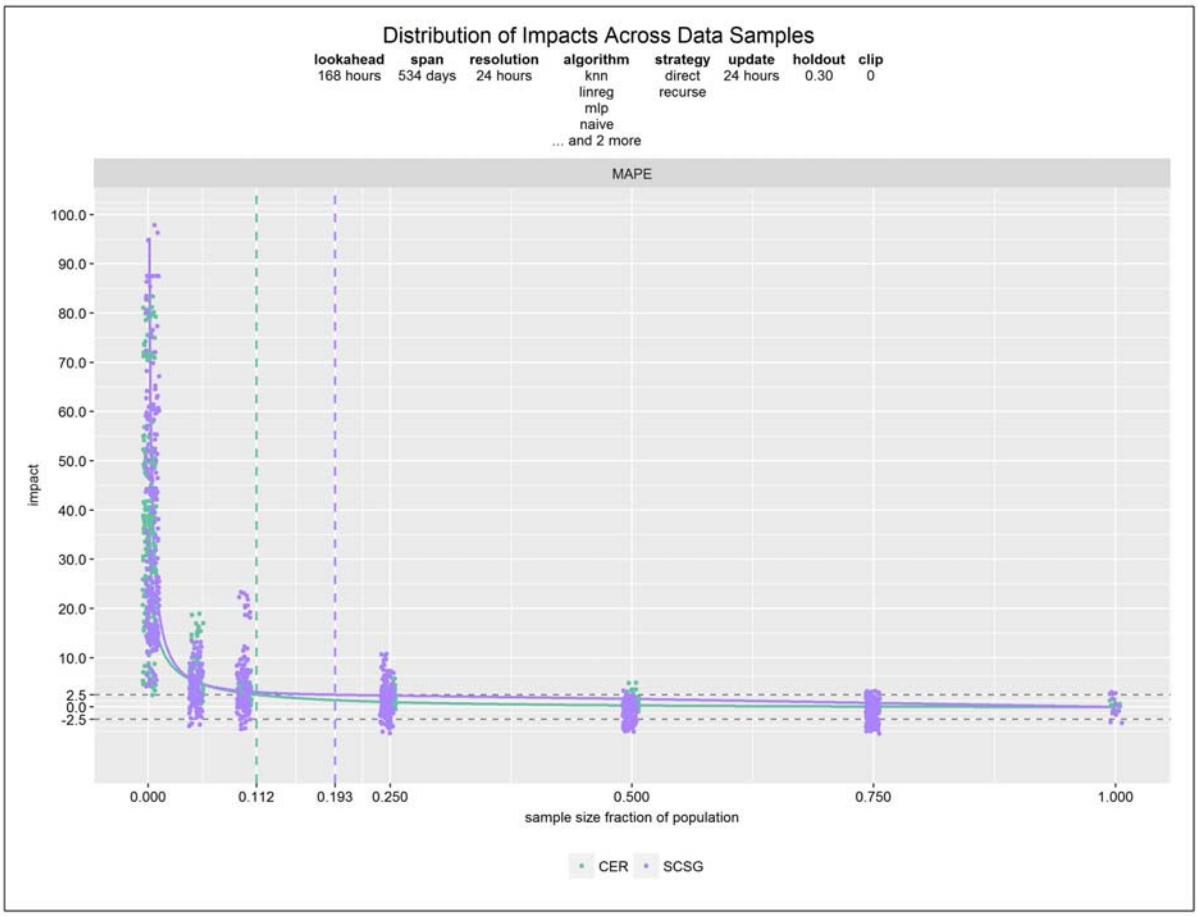


Figure 6-53: Effect of sampling – distribution of impacts vs. sample size, Ireland and Australia, week-ahead forecasts. 12 techniques. For Ireland, 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. For Australia, 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on sample and forecasts based on full population (no sampling). *Green* is Ireland. *Violet* is Australia.

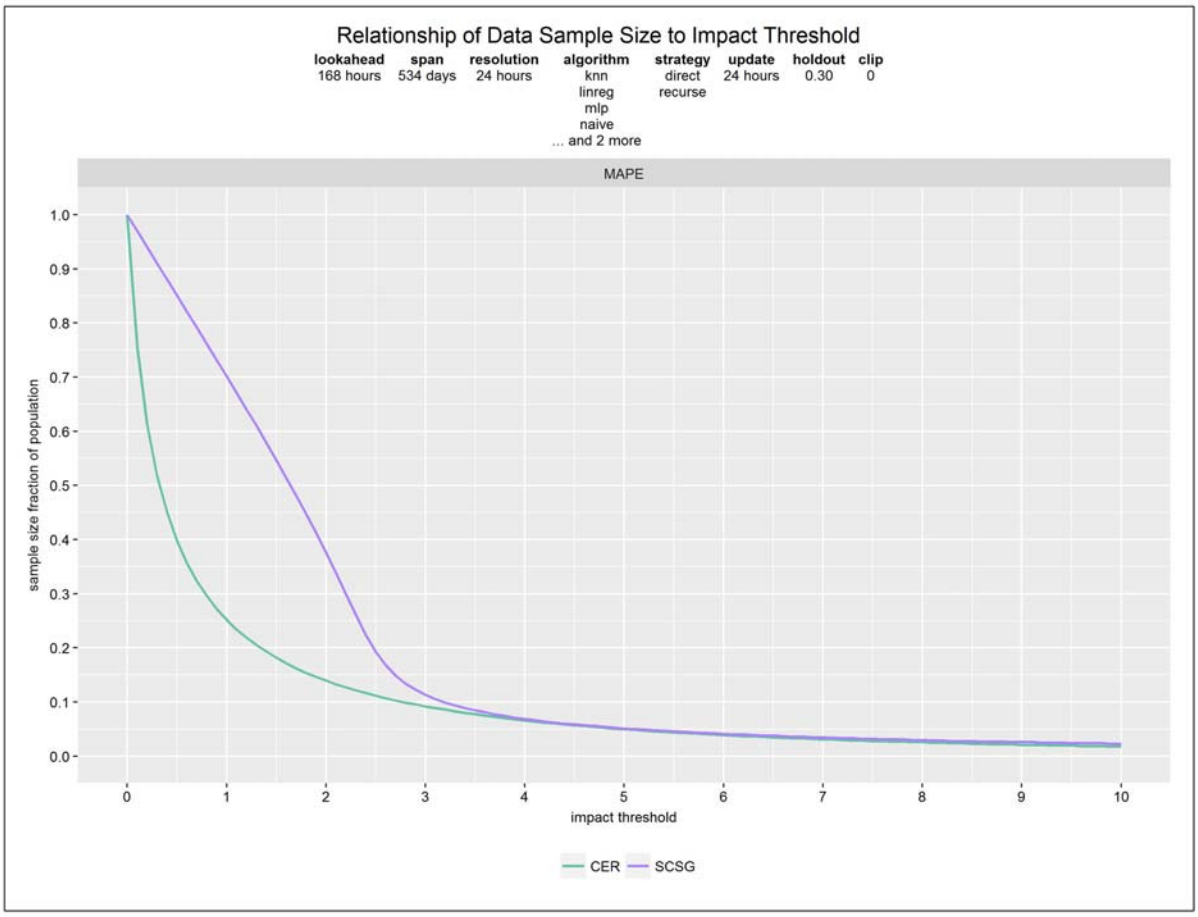


Figure 6-54: Effect of sampling – minimum sample size vs. impact threshold, Ireland and Australia, week-ahead forecasts. 12 techniques. For Ireland, 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. For Australia, 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on sample and forecasts based on full population (no sampling). Impact threshold specifies the upper bound on allowed percentage increase in MAPE. *Green* is Ireland. *Violet* is Australia. For example, when Ireland impact threshold is 100%, minimum sample size is 25% – i.e., if mean MAPE of forecasts on a sample compared to mean MAPE of forecasts on population is not allowed to increase by more than 100%, then the sample size cannot be smaller than 25% of the population size.

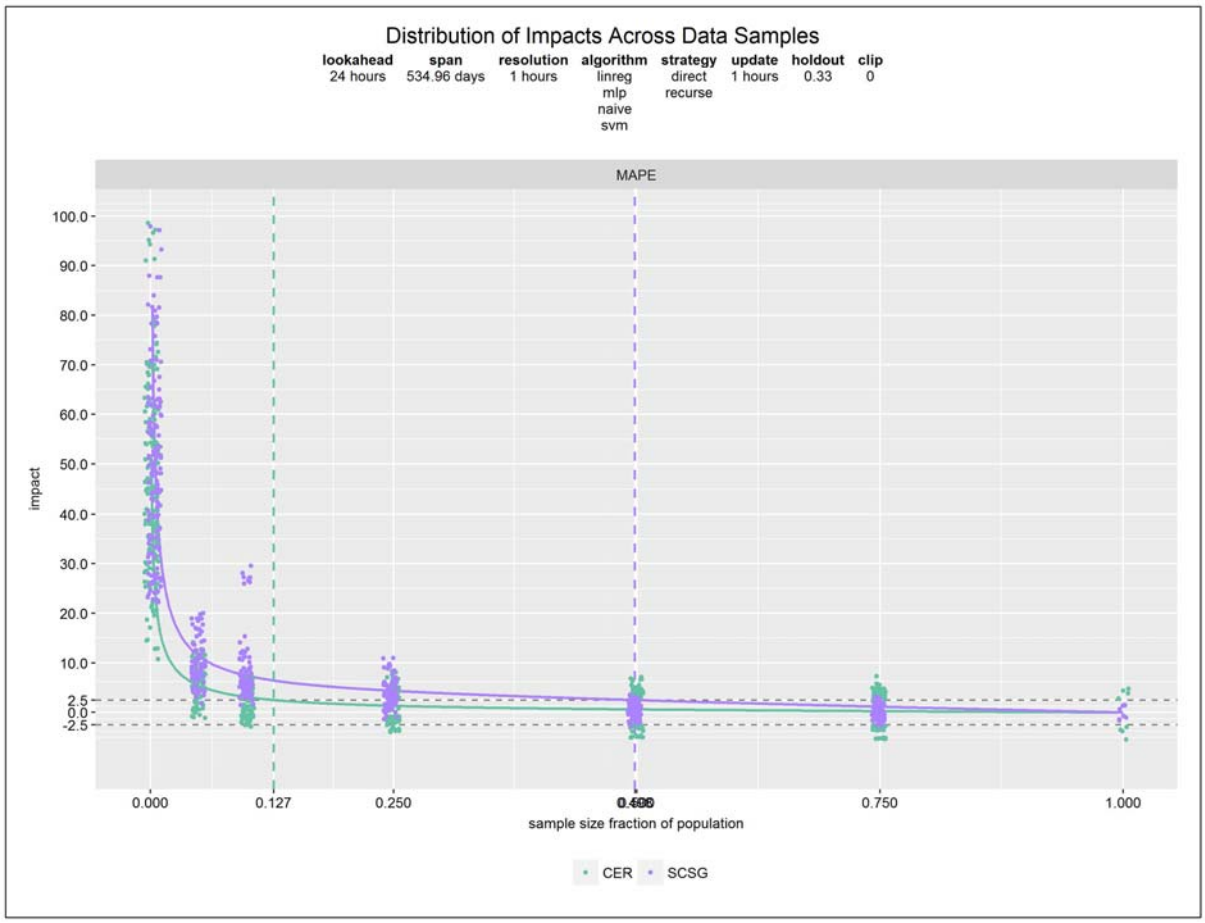


Figure 6-55: Effect of sampling – distribution of impacts vs. sample size, Ireland and Australia, day-ahead forecasts. 8 techniques. For Ireland, 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. For Australia, 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on sample and forecasts based on full population (no sampling). *Green* is Ireland. *Violet* is Australia.

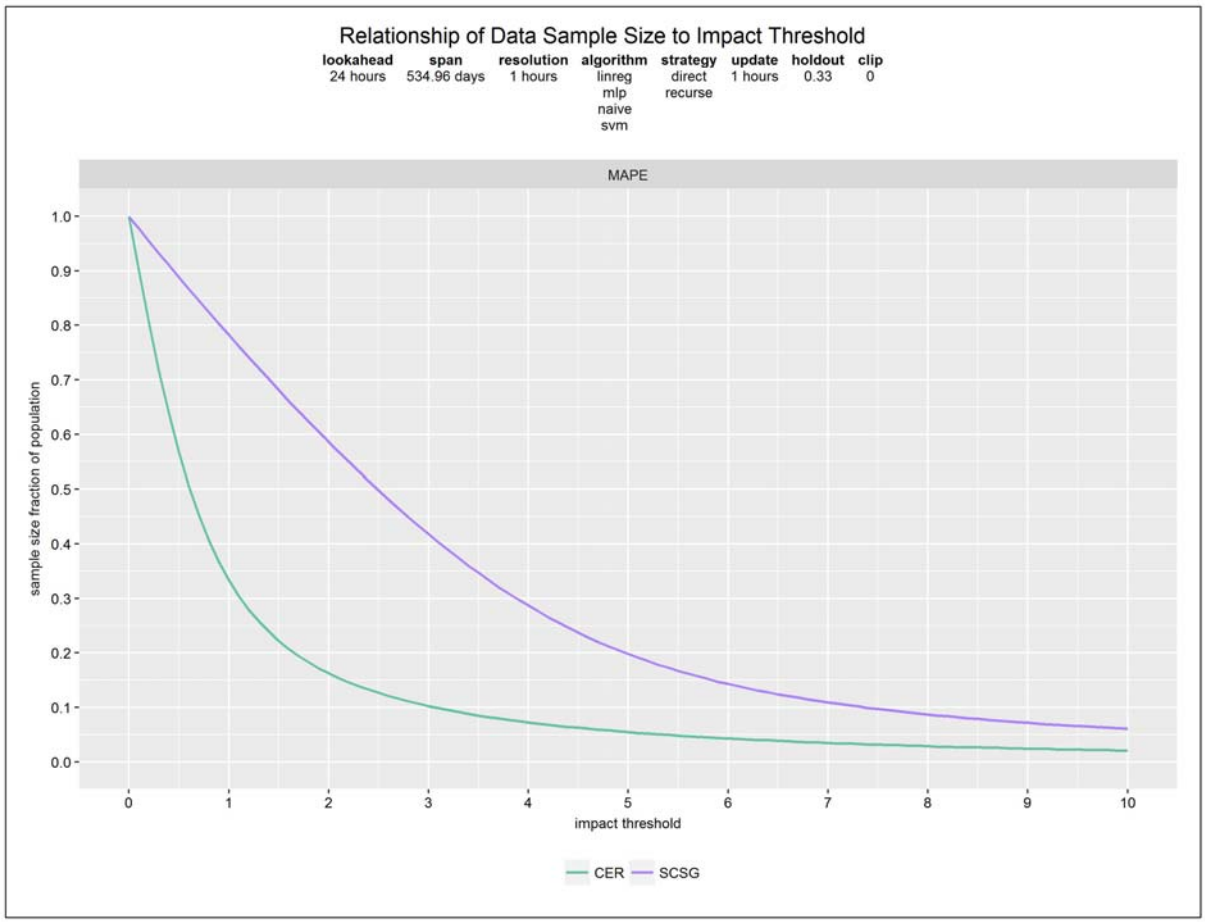


Figure 6-56: Effect of sampling – minimum sample size vs. impact threshold, Ireland and Australia, day-ahead forecasts. 8 techniques. For Ireland, 20 samples at sample size 1 out of 782, 10 samples at other sample sizes. For Australia, 20 samples at sample size 1 out of 223, 10 samples at other sample sizes. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on sample and forecasts based on full population (no sampling). Impact threshold specifies the upper bound on allowed percentage increase in MAPE. *Green* is Ireland. *Violet* is Australia. For example, when Ireland impact threshold is 100%, minimum sample size is 34% – i.e., if mean MAPE of forecasts on a sample compared to mean MAPE of forecasts on population is not allowed to increase by more than 100%, then the sample size cannot be smaller than 34% of the population size.

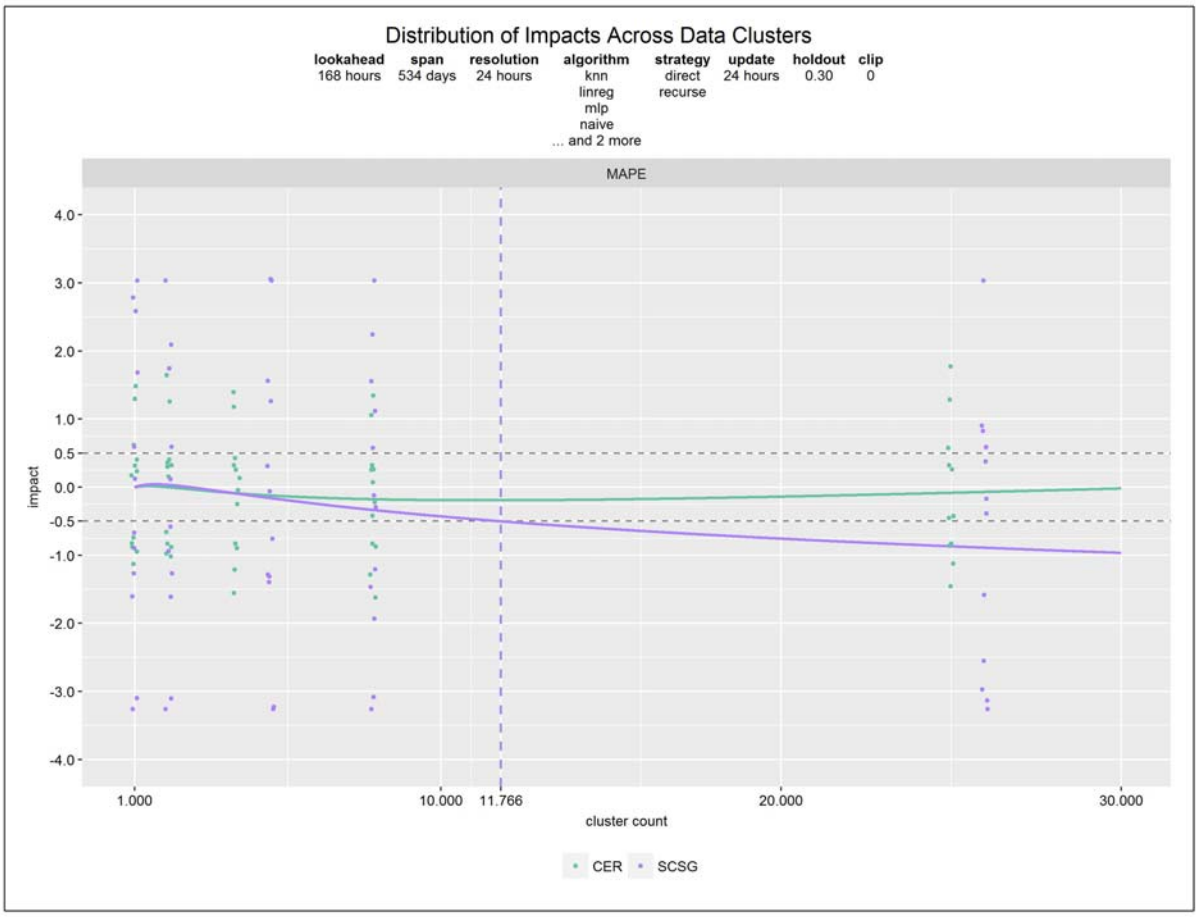


Figure 6-57: Effect of clustering – distribution of impacts vs. cluster count, Ireland and Australia, week-ahead forecasts. 12 techniques. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on clusters and forecasts based on full population (no clustering). *Green* is Ireland. *Violet* is Australia.

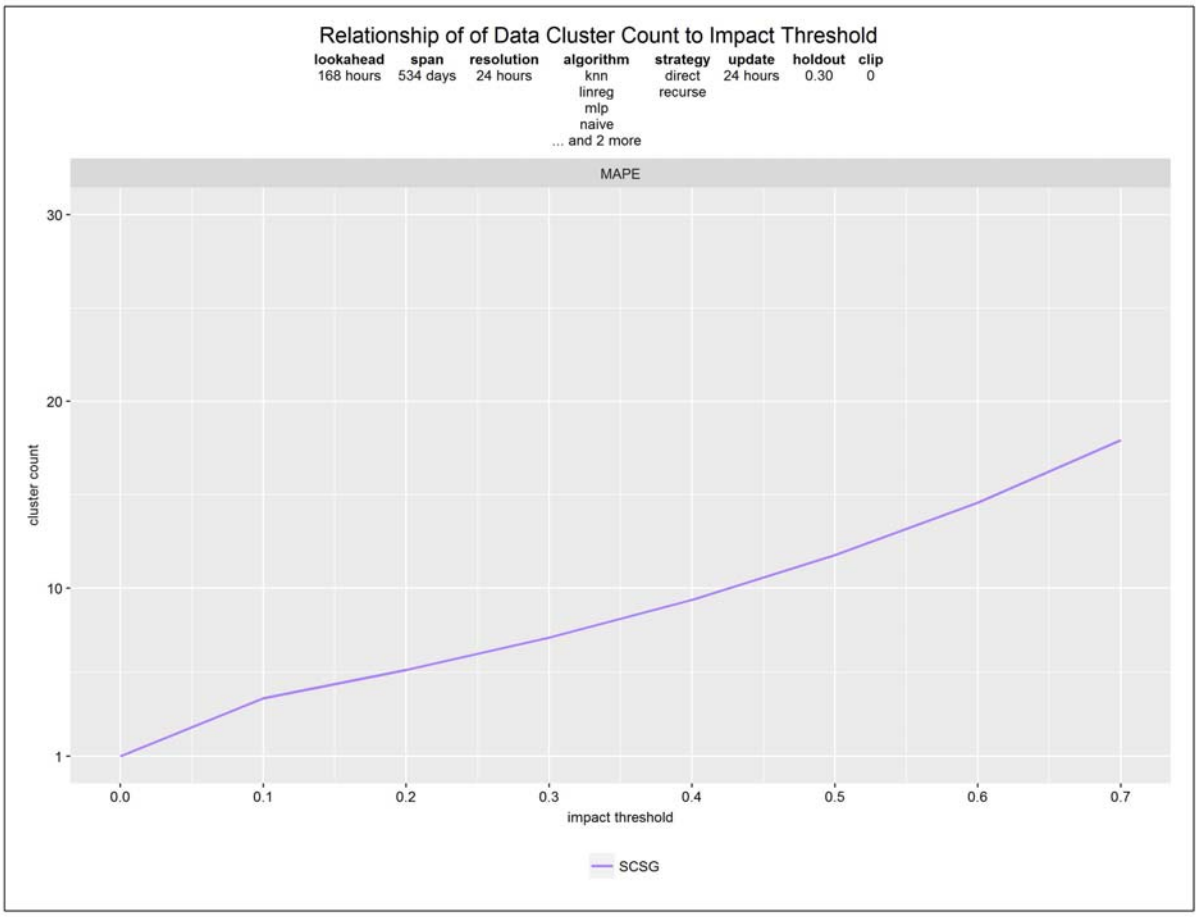


Figure 6-58: Effect of clustering – maximum cluster count vs. impact threshold, Ireland and Australia, week-ahead forecasts. 12 techniques. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on clustering and forecasts based on full population (no clustering). Impact threshold specifies the upper bound on allowed percentage increase in MAPE. *Green* is Ireland. *Violet* is Australia. For example, when Australia impact threshold is 20%, maximum cluster count is 5 – i.e., if mean MAPE of forecasts based on a group of clusters compared to mean MAPE of forecasts on population is not allowed to increase by more than 20%, then the grouping cannot exceed 5 clusters.

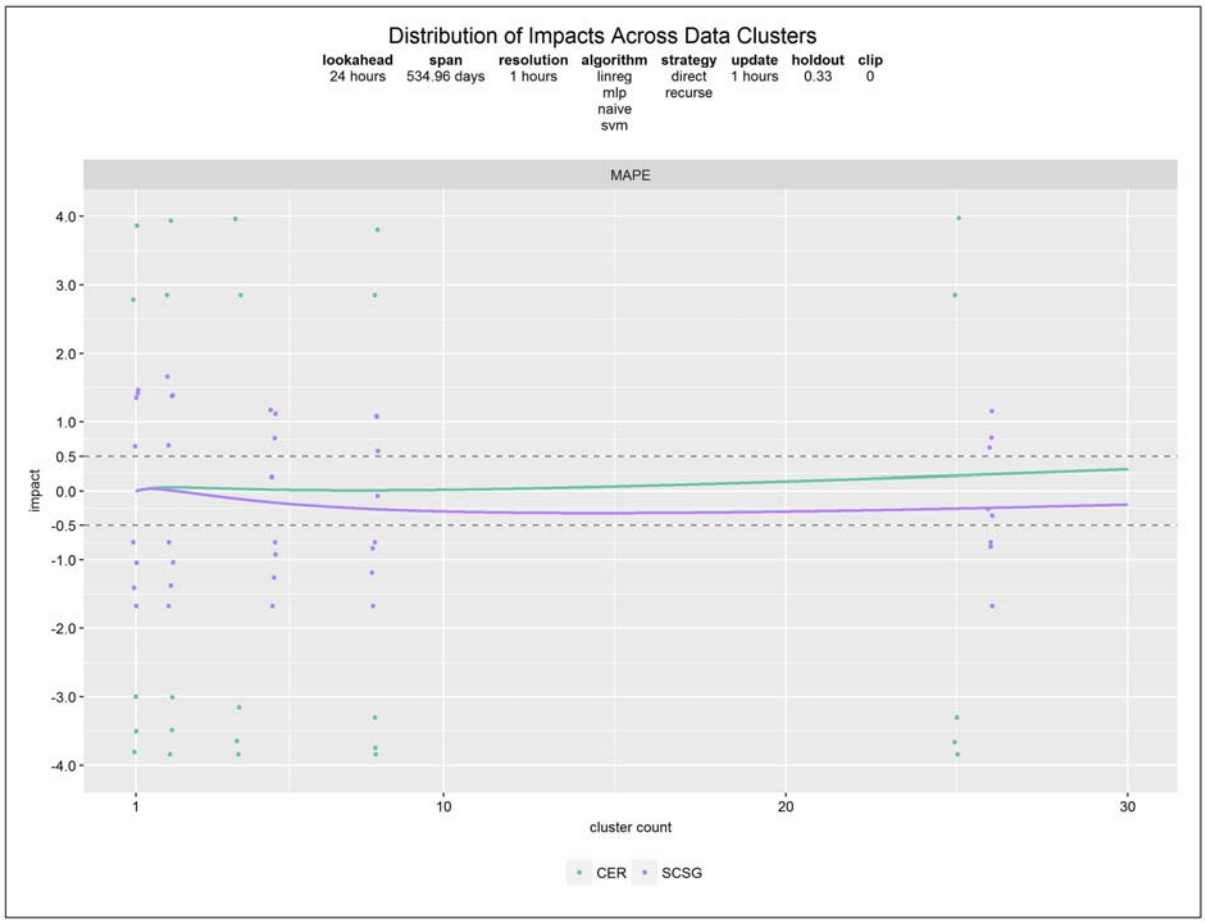


Figure 6-59: Effect of clustering – distribution of impacts vs. cluster count, Ireland and Australia, day-ahead forecasts. 8 techniques. Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on clusters and forecasts based on full population (no clustering). *Green* is Ireland. *Violet* is Australia.

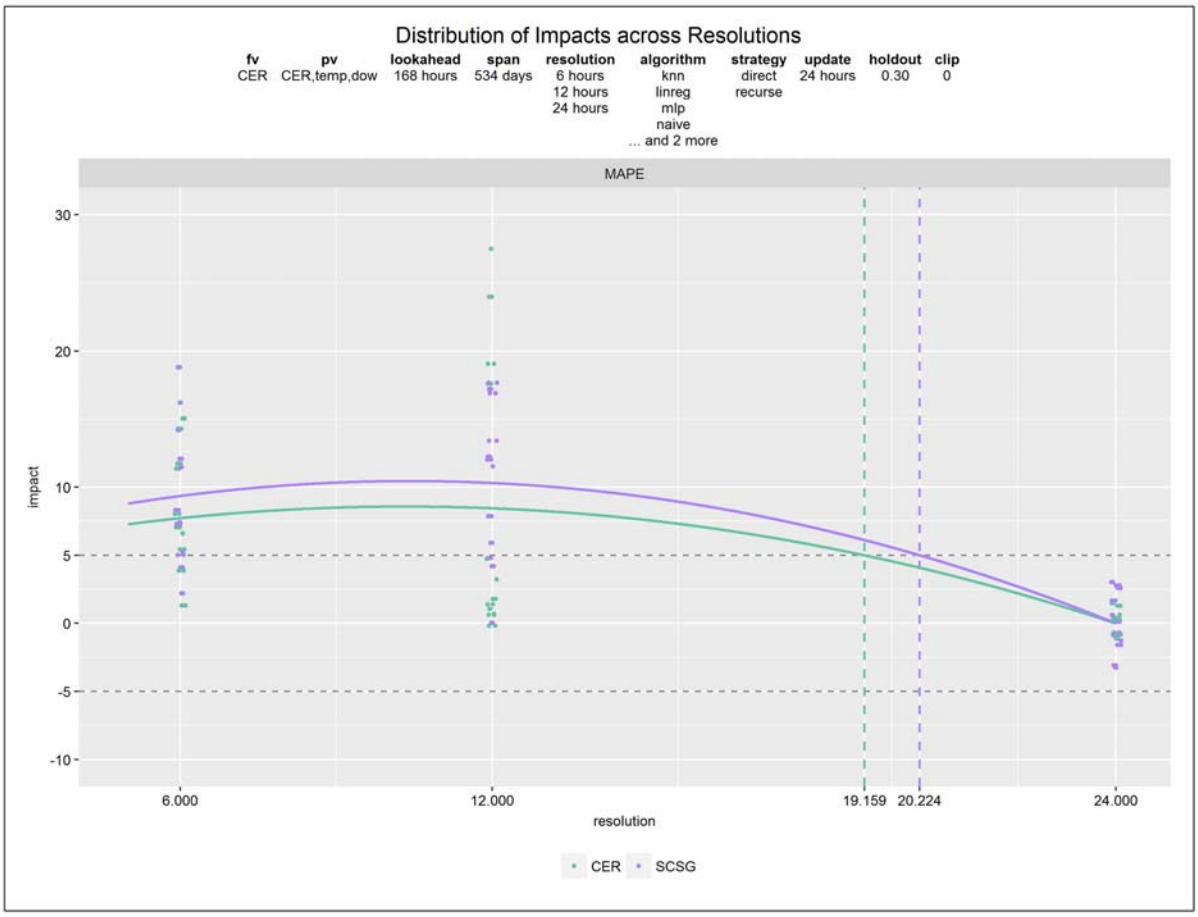


Figure 6-60: Effect of temporal magnification – distribution of impacts vs. time step size, Ireland and Australia, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on temporal magnification and forecasts at a baseline time step size. *Green* is Ireland. *Violet* is Australia.

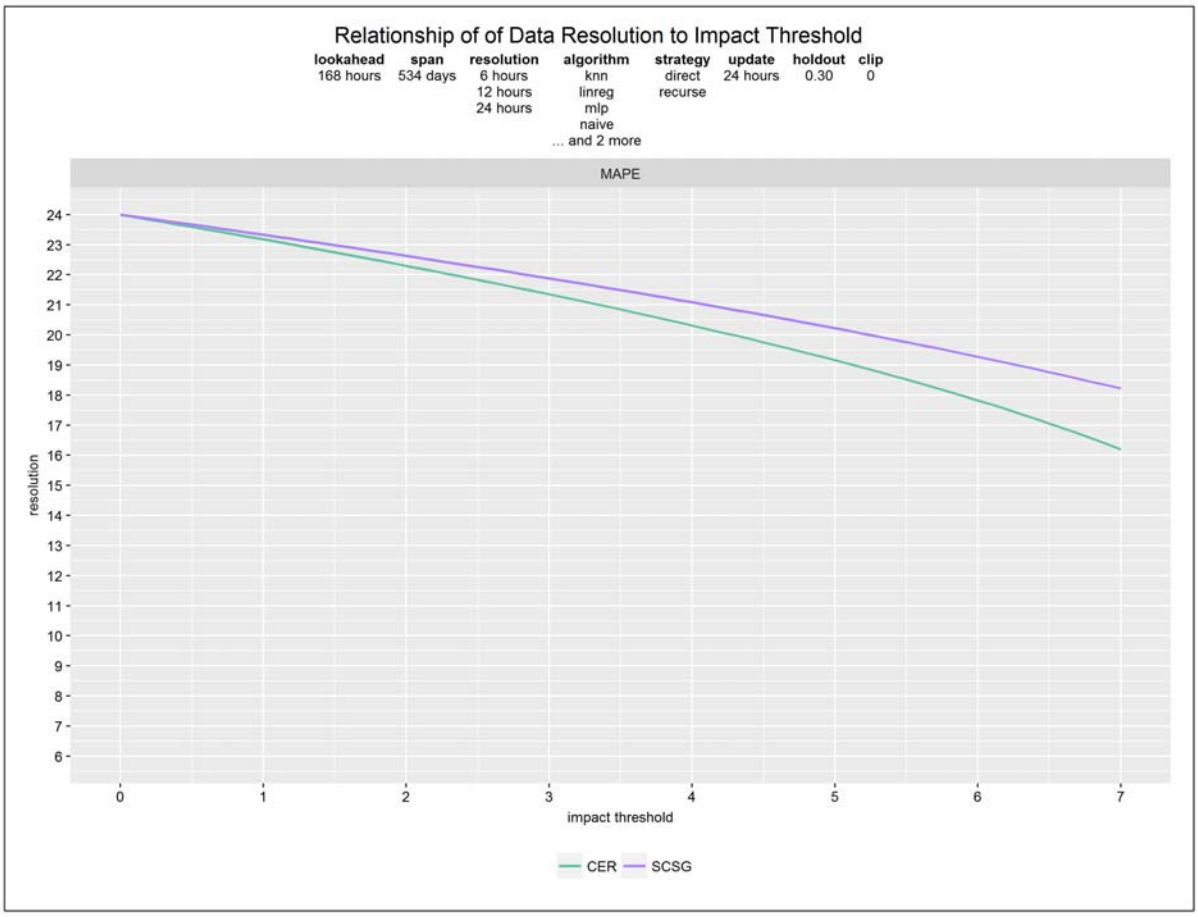


Figure 6-61: Effect of temporal magnification – minimum time step size vs. impact threshold, Ireland and Australia, week-ahead forecasts. 36 techniques (3 decision options for time step size, 12 combinations of other decision options). Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on temporal magnification and forecasts at a baseline time step size. Impact threshold specifies the upper bound on allowed percentage increase in MAPE. *Green* is Ireland. *Violet* is Australia. For example, when Ireland impact threshold is 100%, minimum time step size 23 hours – i.e., if mean MAPE of forecasts based on temporal magnification compared to mean MAPE of forecasts at baseline time step size is not allowed to increase by more than 100%, then the time step size cannot be shorter than 23 hours.

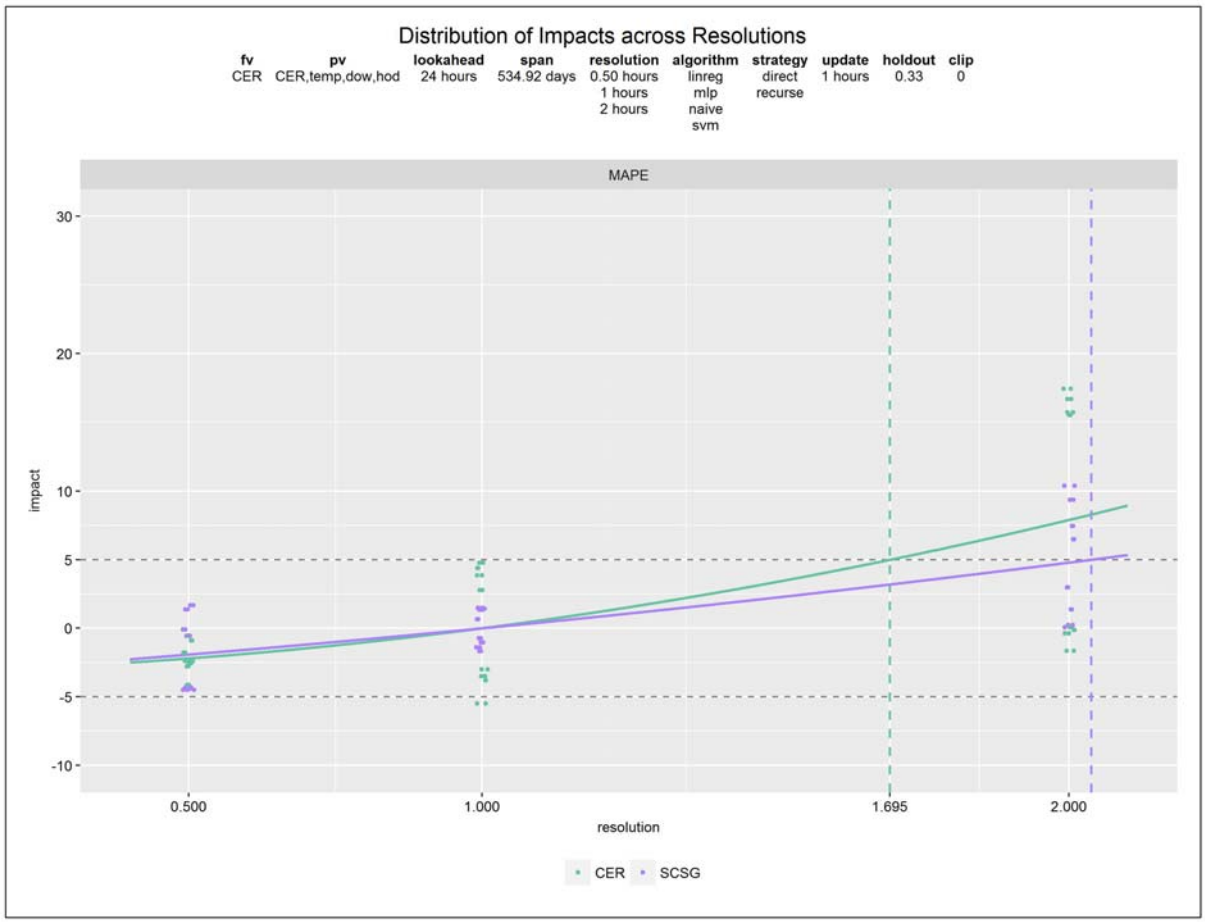


Figure 6-62: Effect of temporal magnification – distribution of impacts vs. time step size, Ireland and Australia, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on temporal magnification and forecasts at a baseline time step size. *Green* is Ireland. *Violet* is Australia.

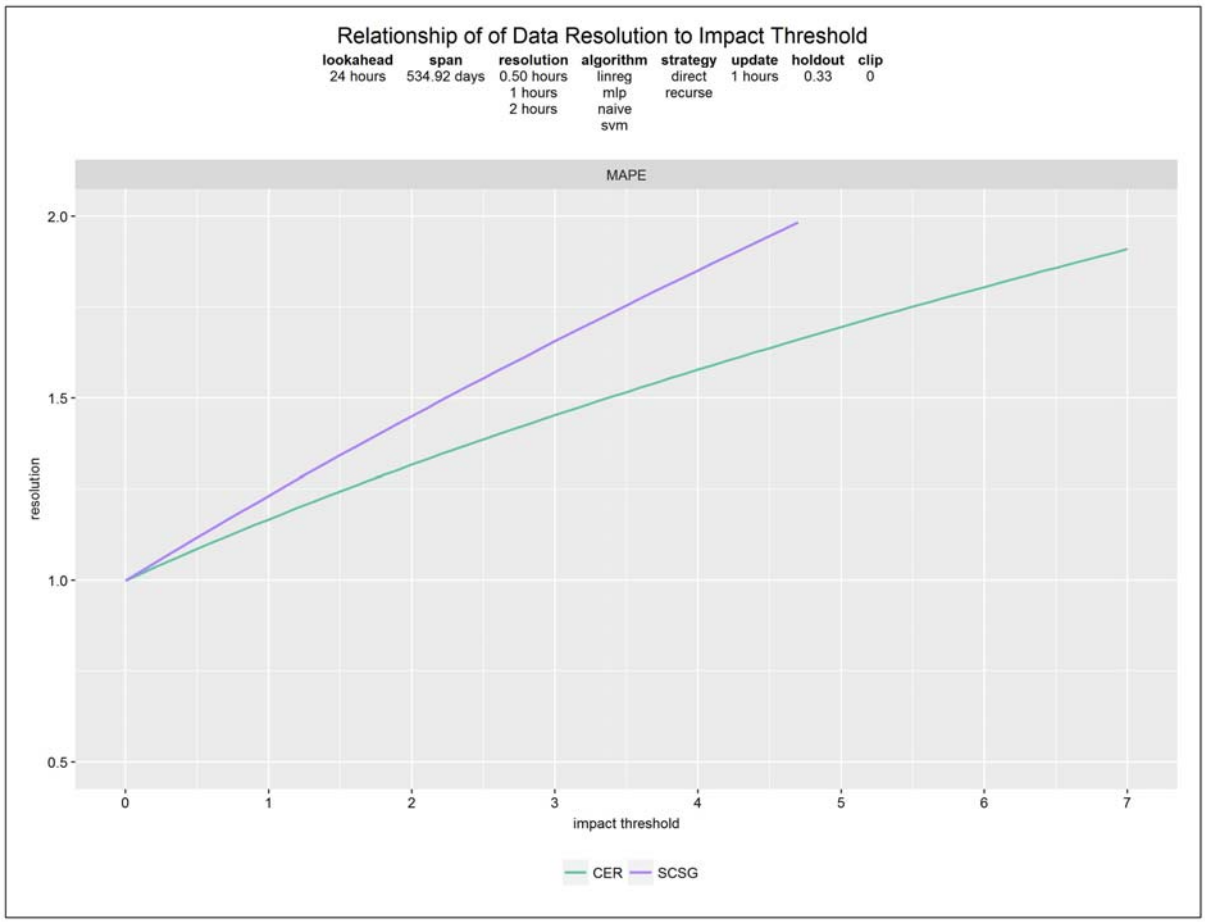


Figure 6-63: Effect of temporal magnification – maximum time step size vs. impact threshold, Ireland and Australia, day-ahead forecasts. 24 techniques (3 decision options for time step size, 8 combinations of other decision options). Metric is MAPE. Impact is defined as percentage difference in MAPE score between forecasts based on temporal magnification and forecasts at a baseline time step size. Impact threshold specifies the upper bound on allowed percentage increase in MAPE. *Green* is Ireland. *Violet* is Australia. For example, when Ireland impact threshold is 100%, minimum time step size 1.2 hours – i.e., if mean MAPE of forecasts based on temporal magnification compared to mean MAPE of forecasts at baseline time step size is not allowed to increase by more than 100%, then the time step size cannot be larger than 1.2 hours.

7 CONCLUSION

"I hope that posterity will judge me kindly, not only as to the things which I have explained, but also to those which I have intentionally omitted so as to leave to others the pleasure of discovery."

– Rene Descartes

We have conducted research into the forecasting process in general, as applied to prediction of forecastability, and as applied to residential electricity demand estimation. Contributions to the field include the following:

Model	A model of the forecasting process that accounts for many of the process decisions studied in the literature.
Analysis Method	A method to analyze the effects of forecasting process decisions about techniques on forecasting performance that accounts for many decisions working in concert.
Analysis Method	A method to analyze the effects of forecasting process decisions about data strategy on forecasting performance that accounts for many decisions working in concert.
Computation Platform	Software to explore the effects of forecasting process decisions.
Proof-of-Concept	Demonstration of the practicality and usefulness of the analysis methods and computation platform applied to benchmark data and real-world data from smart electric grids.
Findings	Several findings about the relationships between forecasting performance and forecasting process decisions, especially in the domain of residential electricity demand estimation.

Our main finding is that forecasting performance is highly sensitive to the combined effects of many forecasting process decisions working in concert. This holds empirically when using M3

Forecasting Competition benchmark datasets, and when using two specific but representative sets of real smart electric grid data from opposite sides of the globe, collected at different times, and reflecting different climates. For the residential electricity demand estimation domain, we see that techniques characterized by too few forecasting process decisions lead to a distorted understanding of their forecasting performance.

We find sampling to be an especially effective data strategy, clustering not so, temporal magnification mixed. Other relationships between certain decisions and performance are surfaced, too.

While these findings are empirical and specific to one practically scoped investigation, they are potentially generalizable, with implications for residential electricity demand estimation, smart electric grid design, and electricity policy.

More research is required to better understand just how far these results generalize. Our analysis methods and computation platform may provide useful guidance and support in this regard. We expect that future studies will increasingly make explicit and account for more forecasting process decisions.

7.1 Summary of Insights

Our analyses reveal several potentially useful insights, with the caveat that they are based on specific datasets and a practically scoped set of experiments.

7.1.1 General Insights

MAIN INSIGHT

- Forecasting performance is sensitive to the interaction effects of many process-level decisions.

OTHER INSIGHTS

- Forecastability is highly sensitive to the metric decision.
- Locked-in decisions can distort the view of forecasting performance.
- Locked-in decisions can distort the view of algorithm class importance to forecasting performance.

7.1.2 Insights About Entropy and Forecastability

MAIN INSIGHT

- Entropy can be a good predictor of forecastability, measured with respect to best performing techniques, and improves as data span lengthens.

OTHER INSIGHTS

- For forecastability defined in terms of a specific technique, entropy is a weak to modest predictor of forecastability, depending on the metric and technique decisions.
- For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, entropy is a strong predictor of forecastability, more or less so depending on the technique decisions for algorithm class and span.
- For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, the predictive power of entropy grows asymptotically along with decisions to increase span.
- For forecastability defined in terms of the set of best performing techniques in the context of many technique decisions, the sensitivity of forecastability to entropy grows along with decisions to increase span.

7.1.3 Insights About Residential Electricity Demand Estimation

MAIN INSIGHTS

- Which process-level decisions are important to forecasting performance are location-specific.
- The best location-specific forecasting techniques do not work well applied to other locations.
- The best “one-size-fits-all” forecasting techniques do work well applied across multiple locations.
- Forecasting based on a small data sample approaches performance of forecasting based on the full population.
- Forecasting based on a cluster grouping of data does not much improve performance over forecasting based on the full population.
- Forecasting based on data collected at relatively fine or coarse resolution may or may not improve performance.

OTHER INSIGHTS

- Relative forecasting performance of algorithms is highly sensitive to the interaction effects of the algorithm decision and other forecasting process decisions.
- Forecasting performance is highly sensitive to the combined effects of forecasting process decisions.
- The best forecasting techniques tend to use a short update cycle.
- Many of the best forecasting techniques use the linear regression algorithm.
- The best week-ahead forecasting techniques use the naïve algorithm when span is short.
- The best week-ahead forecasting techniques use any algorithm except support vector regression when span is long.
- The best day-ahead forecasting techniques use linear regression, direct extension, and short update cycle; other decisions are not as important to performance.
- The best day-ahead forecasting techniques use the naïve or support vector regression algorithm when the update cycle is long.

- Forecasting performance is not much degraded by sampling.
- Forecasting performance is not much improved by clustering.
- Forecasting performance is degraded by refining temporal magnification at coarse time step sizes.
- Forecasting performance is improved by refining temporal magnification at fine time step sizes.

- Training and testing decisions can dominate other forecasting process decisions.
- Different locations lend themselves to different forecasting techniques.
- Forecasting performance at different locations is not much degraded by “one-size-fits-all” techniques.
- Forecasting performance at different locations is not much degraded by sampling.
- Forecasting performance at different locations is not much improved by clustering.

- Forecasting performance at different locations is degraded by refining temporal magnification at coarse time step sizes.

7.2 Future Research

Our analysis results motivate us to pursue more research in several areas.

7.2.1 More on Data Characterization

Our findings about entropy as a predictor of forecastability motivate us to expand the scope of analysis to better quantify and gather more evidence for the relationships between entropy, forecastability, and other factors. While our analysis uses benchmark data, which facilitates comparison of results with other those of other studies, more can be learned by using much longer span data. Also, using a greater variety of algorithm class, metric, and other decision options will provide more support for whatever relationships are found. Further, the same investigative approach can be applied to look for predictive relationships between a wide variety of data characteristics and the importance of various decisions. Such predictive relationships, if known, can potentially be used to inform choice of clustering strategies, so that individual clusters are optimized for maximum forecastability. All of these investigations constitute finding "a predictor of a predictor", so machine learning can be leveraged to find both predictors.

- More on entropy as a predictor of forecastability – longer time series, multi-variate algorithms, variety of metrics
- Data characterization as a predictor of decision importance
- Cluster similarity measure as a predictor of decision importance
- Machine learning for forecastability prediction models
- Machine learning for data importance prediction models

7.2.2 Expand Scope of Analysis on Residential Electricity Demand Estimation

To quantify and verify the robustness of the relationships we find between decisions and forecasting performance, we look to expand the scope of our analysis in several dimensions, including especially repeating the analyses with data sourced from additional locations.

- More decisions, decision options, metrics, reference series, predictor series
- Verify robustness of relationships, quantify relationships

7.2.3 Formulate the Cost Function for Residential Electricity Demand Estimation

As discussed, we expect that penalty functions correspond closely to real economic costs, which would make them a practical tool for forecasting practitioners responsible for the economic cost implications of their forecasts. The form, variables, range of variable values, and ties to penalty functions must be gathered from an exploration of real forecasting practitioner environments, like utility company planning department operations.

- Identify variables and variable value ranges for cost function

7.2.4 Expand Scope of Analysis to Other Domains

In addition to analyses presented in this dissertation, we have started applying our methods and platform to analyze forecasting performance in other domains, specifically forecasts of technical support call center incoming call levels and international refugee levels. For calls, we have been provided long-term, fine-resolution data, and so find that the methods work well. For refugees, available data is much coarser, and so using the methods in this domain produces less satisfying results.

- Technical support call centers
- International refugees

- Non-residential electricity demand estimation
- Other domains

7.2.5 Expand Computation Platform Functionality

- Large scale data handling
- Memory & speed optimization
- Standardized data pre-processing
- More built-in algorithm classes and metrics
- Complex algorithm tuning
- Complex integration rules
- Variable decisions per sample or cluster
- Complex training and testing rules
- Simplified architecture and data organization
- More visualizations

7.2.6 Prepare Computation Platform for Commercial or Open Source Use

Full proof-of-capability of the computation platform, and readiness for use by a wide audience, requires further enhancements. One approach is to structure it as an R library of functions to be distributed as open source software. Another approach is to wrap it in a user interface, perhaps implemented with the Shiny R library, to make its functionality available to non-programmers.

- R library
- Shiny user interface
- User documentation

Appendix A MORE ABOUT SAMPLING RULES

A.1 Bootstrap Sampling, version 1

[Low level of control over randomness in resulting sample] Start with a population of n observations. Pick n-1 observations at random with replacement. The sample will be the same size as the original population and some of the observations will be duplicates. Repeat for as many samples as desired. This approach does not control how much of the original population is reflected in the sample, nor how much influence particular observations have on the sample.

A.2 Bootstrap Sampling, version 2

[Medium-low level of control over randomness in resulting sample] Start with a population of n observations. For some i in (1 through n-1), pick i observations at random without replacement. From this set, pick n observations at random with replacement. The sample will be the same size as the original population and some of the observations will be duplicates. Repeat for as many samples at i as desired. Repeat for each i as desired. This approach controls how much of the original population is reflected in the sample, but does not control how much influence particular observations have on the sample.

A.3 Bootstrap Sampling, version 3

[Medium-high level of control over randomness in resulting sample] Start with a population of n observations. For some i in (1 through n-1), pick i observations without replacement. From this set, duplicate each observation an approximately equal number of times until reaching exactly n observations. The sample will be the same size as the original population and some of the observations will be duplicates. Repeat for as many samples at i as desired. Repeat for each i as desired. This approach controls how much of the original population is reflected in the sample, and ensures each observation has approximately equal influence on the sample.

A.4 Jackknife Sampling

[High level of control over randomness in resulting sample] Start with a population of n observations. For some i in (1 through n-1), pick i observations without replacement. From this set, scale each observation by n/i. Repeat for as many samples at i as desired. Repeat for each i as desired. The sample will be smaller than the original population, but any aggregations applied to the original population and sample will result in comparable magnitudes. This approach controls how much of the original population is reflected in the sample, and ensures each of these observations has exactly equal influence on the sample.

Appendix B MORE ABOUT EXTENSION RULES

B.1 Direct Extension Rules

At training time, several engines are produced based on the training period in advance of any forecasting in the test period. All the engines accept as input a set of predictor series levels at time steps prescribed by the look-back, but each specific engine outputs a forecast for one specific time step related to the look-ahead. The first engine produces a forecast for the time step at the look-ahead, the second engine produces a forecast for one time step beyond that, and so on for all time steps prescribed by the update cycle. At test time, all the engines are applied to the same origin time step using the same inputs, resulting in a forecast starting at a distance from the origin prescribed by the look-ahead and continuing throughout the update cycle. Then, the origin is repeatedly advanced as prescribed by the update cycle, and the engines each time applied again. The forecasts (each covering a period prescribed by the update cycle) are concatenated to produce a forecast covering the complete test period.

Figure B-1 illustrates an example. The look-ahead is 3 time steps, the update cycle is 2 time steps, and the look-back is the set of 0, 1, 2, 3, 4, 5 time steps. At training time, 2 engines are produced: the first forecasts 3 time steps ahead, the second forecasts 4 time steps ahead. At test time, both engines are applied to the same origin time step, indicated in blue, using the same inputs, namely the predictor series levels at 0, 1, 2, 3, 4, and 5 steps behind the origin. The first engine outputs a 3 time step ahead forecast, indicated as "?" (a). The second engine outputs a 4 time step ahead forecast, also indicated as "?" (b). The origin is then advanced 2 time steps and the procedure repeated throughout the test period (c).

B.2 Recursive Extension Rules

At training time, one engine is produced based on the training period in advance of any forecasting in the test period. The engine accepts as input a set of predictor series levels at time steps prescribed by the look-back, and outputs a forecast for one time step ahead,

regardless of the look-ahead. At test time, the engine is applied to the origin time step, resulting in a forecast for one time step ahead of the origin. Subsequently, the origin is advanced one time step and the engine applied again, this time with new input potentially including some of the recent forecasts. This is repeated until forecasts are produced for time steps starting at a distance from the origin prescribed by the look-ahead and continuing throughout the update cycle. Then, the origin is repeatedly advanced as prescribed by the update cycle, and the procedure applied again. The forecasts (each covering a period prescribed by the update cycle) are concatenated to produce a forecast covering the complete test period.

Figure B-2 illustrates an example. The look-ahead is 3 time steps, the update cycle is 2 time steps, and the look-back is the set of 0, 1, 2, 3, 4, 5 time steps. At training time, one engine is produced that forecasts 1 time step ahead. At test time, the engine is applied to the origin time step, indicated in blue, using as inputs the predictor series levels at 0, 1, 2, 3, 4, and 5 steps behind the origin. It outputs a 1 time step ahead forecast, indicated as "?" (a). Subsequently, the origin is advanced 1 time step and the engine applied again, using new inputs accordingly, including the forecast just made (b). This is repeated until forecasts have been produced for each time step covered by the update cycle (c) (d). The origin is then advanced 2 time steps and the procedure repeated throughout the test period (e).

B.3 Other Extension Rules

A variation on either the direct or recursive rule is to periodically produce new engines with refreshed training data as prescribed by the retrain cycle. If practical to do so, this would potentially guard against the engines losing relevance near the end of the test period. Many other variations have been proposed in the literature.

Note, in the degenerate case of look-ahead 1 time step, update cycle 1 step, i.e., continuously advancing the origin, the direct and recursive rules produce the same forecast.

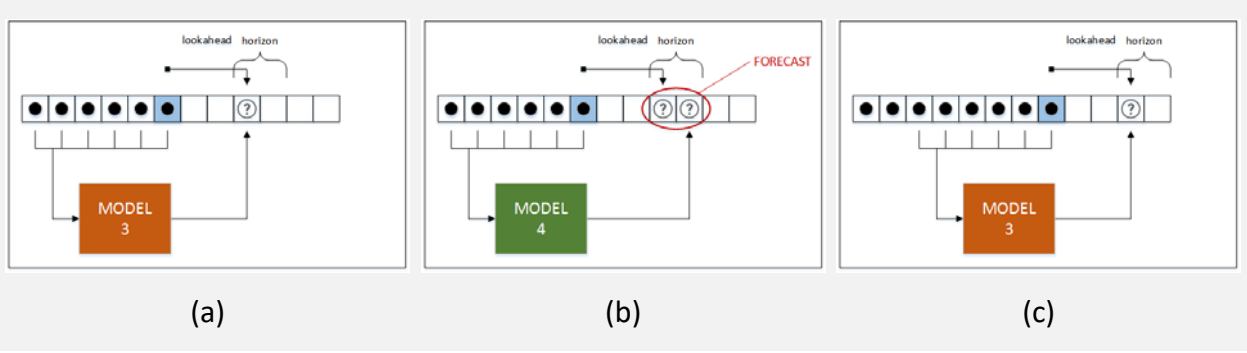


Figure B-1: Direct extension rule, look-ahead=3, update cycle=2, look-back=(0,1,2,3,4,5)

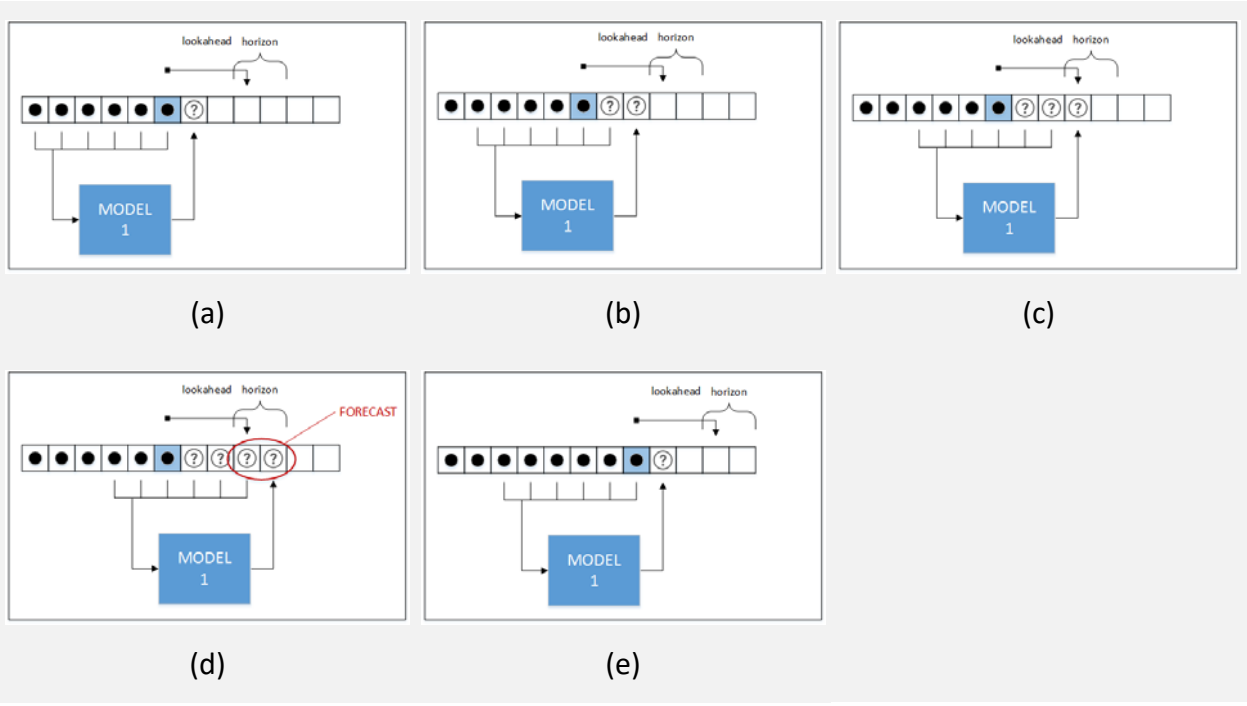


Figure B-2: Recursive extension rule, look-ahead=3, update cycle=2

REFERENCES

Overview of Forecasting

1. Adhikari R, Agrawal R. *An Introductory Study on Time Series Modeling and Forecasting.* Germany: Lambert Academic Publishing; 2013.

   --- *Some popular time series forecasting models used in practice. Deescribes three important classes of time series models: the stochastic, neural networks, and SVM based models. Stationarity, parsimony, overfitting, etc. Five performance measures: MSE, MAD, RMSE, MAPE, Theil's U-statistics. For each of six datasets, show the obtained forecast diagram.*

2. De Gooijer JG, Hyndman RJ. 25 years of time series forecasting. *International Journal of Forecasting.* 2006; 22(3):443-473.

   --- *Review of 25 years of research into time series forecasting, 1982–2005.*

3. Diebold FX. *Elements of Forecasting.* 4 ed: South-Western College Pub; 2006.

   --- *Textbook covering forecasting principles, including a particular model of the forecasting process.*

4. Hyndman RJ. Use fake data and real data. *Hyndsight.* 2010 (Jun 11, 2010). http://robjhyndman.com/hyndsight/use-fake-data-and-real-data/.

   --- *Advantages/disadvantages of testing new forecasting techniques on synthetic vs. real data.*

5. Hyndman RJ. Benchmarks for Forecasting. *Hyndsight.* 2010 (Aug 25, 2010). http://robjhyndman.com/hyndsight/benchmarks/.

   --- *Importance of testing new forecasting techniques against benchmark techniques.*

6. Hyndman RJ, Athanasopoulos G. *Forecasting: Principles and Practice.* OTexts; 2013.

   --- *Textbook covering forecasting principles. Good introduction to forecasting.*

7. Lorenz R, Dannecker L, Rösch P, Lehner W, Hackenbroich G, Schlegel B. Forecasting in Hierarchical Environments. Paper presented at: International Conference on Scientific and Statistical Database Management (2012).

--- *Novel hierarchical forecasting approach, push forecast models to the entities on the lowest hierarch level and reuse these models to efficiently create forecast models on higher hierarchical levels.*

8. Shumway RH, Stoffer DS. *Time Series Analysis and Its Applications: with R Examples.* 3 ed: Springer; 2011.

--- *Textbook covering time series analysis. Relatively advanced material. Includes R examples.*

Overview of Linear Models

9. Dawes RM, Corrigan B. Linear models in decision making. *Psychological Bulletin.* 1974; 81(2):95-106.

--- *The classic seminal article on linear models. Review contexts, shows common structural characteristics. Four examples involving the prediction of codable output variables as grade point average and psychiatric diagnosis are analyzed in detail. In all four, random linear models yield predictions that are superior to those of human judges.*

10. Hogarth RM, Karelaia N. Heuristic and linear models of judgment: matching rules and environments. *Psychological Review.* 2007; 114(3):733-758.

--- *Implications of judgmental heuristics. Judgment modeled in the form of as if linear models. Uses statistical tools to model how the performance of heuristic rules varies as a function of environmental characteristics.*

Overview of Machine Learning

11. Chapman P, Clinton J, Kerber R, et al. *CRISP-DM 1.0.* SPSS; (2000).

--- *The CRISP-DM methodology is described in terms of a hierarchical process model, comprising sets of tasks described at four levels of abstraction: phase, generic task, specialized task, and process instance.*

12. Wirth R, Hipp J. CRISP-DM: Towards a standard process model for data mining. Paper presented at: International Conference on the Practical Application of Knowledge Discovery and Data Mining (2000).

--- *The CRISP-DM (CRoss Industry Standard Process for Data Mining) project proposed a comprehensive process model for carrying out data mining projects.*

13. Witten IH, Frank E, Hall MA. *Data Mining: Practical Machine Learning Tools and Techniques.* 3 ed: Morgan Kaufmann; 2011.

--- *Textbook covering a variety of machine learning algorithms.*

Overview of Electricity Demand Forecasting

14. Alfares HK, Nazeeruddin M. Electric load forecasting: Literature survey and classification of methods. *International Journal of Systems Science.* 2002; 33(1):23-34.

--- *A review and categorization of electric load forecasting techniques. Techniques classified into nine categories: (1) multiple regression, (2) exponential smoothing, (3) iterative reweighted least-squares, (4) adaptive load forecasting, (5) stochastic time series, (6) ARMAX models based on genetic algorithms, (7) fuzzy logic, (8) neural networks and (9) expert systems.*

15. Almeshaiei E, Soltan H. A methodology for Electric Power Load Forecasting. *Alexandria Engineering Journal.* 2011; 50(2):137-144.

--- *A pragmatic methodology that can be used as a guide to construct electric power load forecasting models. Based on decomposition and segmentation of the load time series. Several statistical analyses are involved to study the load features and forecasting precision such as moving average and probability plots of load noise. Real daily load data from Kuwaiti electric network are used as a case study.*

16. Campbell PR, Adamson K. Methodologies for Load Forecasting. Paper presented at: International Conference on Intelligent Systems (2006).

17. Hong T. Energy forecasting: past, present and future. *Foresight: The International Journal of Applied Forecasting.* 2014; 2014(32):43-48.

--- *Practical overview of energy forecasting.*

18. Hong T. *Short Term Electric Load Forecasting* [PhD Thesis]: Operations Research, North Carolina State University; 2010.

--- *A clear, useful overview of electricty demand forecasting. Proposal for an integrated forecasting framework with the concentration on the short term load forecasting (STLF) engine that can easily link to various other forecasts. Disassembles the major techniques that have been applied to STLF and reported in the literature, and reassembles the key elements to come up with a methodology to analyze STLF problems and develop STLF models. Multiple linear regression (MLR) analysis is deployed in the case study of a US utility.*

19. Hong T, Fan S. Probabilistic electric load forecasting: A tutorial review. *International Journal of Forecasting.* 2016.

--- *Tutorial review of probabilistic electric load forecasting, including notable techniques, methodologies and evaluation methods, and common misunderstandings.*

20. Hong T, Laing TD, Pu W. Four best practices of load forecasting for electric cooperatives. Paper presented at: 2014 IEEE Rural Electric Power Conference (REPC); May 18-21, 2014 (2014).

21. Hong T, Shahidehpour M. *Load Forecasting Case Study.* Eastern Interconnection States' Plannig Section; Jan 15, 2015 (2015).

--- *University of North Carolina at Charlotte (UNCC) teamed with Illinois Institute of Technology (IIT), ISO-New England, and North Carolina Electric Membership Corporation (NCEMC) to prepare a Load*

Forecasting Case Study for the Eastern Interconnection States' Planning Council (EISPC) in response to the NARUC solicitation.

22. Hyndman RJ, Fan S. Monash Electricity Forecasting Model May 28, 2015.

--- *The Monash Electricity Forecasting Model (MEFM) was developed to forecast the probability distribution of electricity demand, and the underlying theory and methodology. The model has been used to forecast the probability distribution of annual, seasonal and weekly peak electricity demand and energy consumption for various regions of Australia, including the regions of the National Electricity Market (NEM), the SWIS of Western Australia and the Keilor Terminal Station in Victoria. Short-term forecasting software has also been developed to forecast half-hourly electricity demand (up to seven days ahead) in South Australia and Victoria.*

23. Kyriakides E, Polycarpou MM. Short Term Electric Load Forecasting: A Tutorial. *Trends in Neural Computation*: Springer; 2006:391-418.

--- *Tutorial introduction to the short term load forecasting problem and a brief summary of the various approaches that have been proposed, from conventional to computational intelligence methods.*

24. Lai S-H, Hong T. *When one size no longer fits all: electric load forecasting with a geographic hierarchy.* SAS; (2013).

--- *An electric load-forecasting case study. Weather station data can improve the predictive analytics used to determine future electric usage. Retrain models multiple times each year.*

25. Metaxiotis K, Kagiannas A, Askounis D, Psarras J. Artificial intelligence in short term electric load forecasting: a state-of-the-art survey for the researcher. *Energy Conversion and Management.* 2003; 44(9):1525-1534.

--- *An overview of AI technologies, and their current use in the field of short term electric load forecasting (STELF).*

26. Singh AK, Chaturvedi DK. An Overview of Electricity Demand Forecasting Techniques. Paper presented at: National Conference on Emerging Trends in Electrical, Instrumentation & Communication Engineering (2013).

27. Wijaya TK, Sinn M, Chen B. Forecasting Uncertainty in Electricity Demand. Paper presented at: AAAI Conference on Artificial Intelligence (2015); Austin, TX.

--- *Proposal for a novel approach to estimate the time-varying conditional variance of the GAM residuals, which we call the GAM2 algorithm. It allows utility companies and network operators to assess the uncertainty of future electricity demand and incorporate it into their planning processes. The basic idea of our algorithm is to apply another GAM to the squared residuals to explain the dependence of uncertainty on exogenous variables. Combine modeling approach with online learning algorithms that adjust for dynamic changes in the distributions of demand.*

Machine Learning Algorithms

28. Collobert R, Bengio S. Links Between Perceptrons, MLPS and SVMs. Paper presented at: International Conference on Machine Learning (2004); Banff, Canada.

--- *Links between three important classification algorithms: Perceptrons, Multi-Layer Perceptrons (MLPs) and Support Vector Machines (SVMs).*

29. Hsu C-W, Chang C-C, Lin C-J. *A Practical Guide to Support Vector Classification.* National Taiwan University; updated May 19, 2016; initial version 2003 (2016).

--- *The support vector machine (SVM). Simple procedure which usually gives reasonable results.*

30. Huang G-B, Zhu Q-Y, Siew C-K. Extreme learning machine: Theory and applications. *Neurocomputing.* 2006; 70(1-3):489-501.

--- *Proposal for a new learning algorithm called extreme learning machine (ELM) for single-hidden layer feedforward neural networks (SLFNs) which randomly chooses hidden nodes and analytically determines the output weights of SLFNs. In theory, this algorithm tends to provide good generalization performance at extremely fast learning speed.*

31. Sathyanarayana S. A Gentle Introduction to Backpropagation: a manuscript. Jul 22, 2014 (2014), Numeric Insight.

32. Cortez P. Sensitivity analysis for time lag selection to forecast seasonal time series using Neural Networks and Support Vector Machines. Paper presented at: International Joint Conference on Neural Networks (IJCNN) (2010); Barcelona.

--- *Simultaneous variable (i.e. time lag) and model selection algorithm for multi-step ahead forecasting using NN and SVM. Variable selection is based on a backward algorithm that is guided by a sensitivity analysis procedure, while model selection is achieved using a grid-search. Several experiments were devised by considering eight seasonal series and the forecasts were analyzed using two error criteria (i.e. SMAPE and MSE).*

33. Dietterich TG. Machine Learning for Sequential Data: A Review. In: Caelli T, Amin A, Duin RPW, Kamel M, Ridder D, eds. *Structural, Syntactic, and Statistical Pattern Recogition*: Springer-Verlag Berlin Heidenberg; 2002.

--- *Formalizes the principal learning tasks and describes the methods that have been developed within the machine learning research community. Sliding window methods, recurrent sliding windows, hiddenMarkov models, conditional random fields, and graph transformer networks.*

34. Jiang QY. Time Series Prediction Based on Machine Learning. Paper presented at: International Conference on Electrical, Automation and Mechanical Engineering (EAME) (2015).

--- *Key technologies for predicting the time series by machine learning: (1) a two-class nuclear space feature selection frame (2), a fast multi-core-based distance learning method (3) the prediction of time series based on key technologies applied to machine learning model among business intelligence.*

35. Khadka M, Popp B, George KM, Park N. A New Approach for Time Series Forecasting based on Genetic Algorithm. Paper presented at: International Conference on Computer Applications in Industry and Engineering (CAINE) (2010); Las Vegas, Nevada.

--- *Proposal for a new fusion approach to predict time series based on Concordance and Genetic Algorithm. Different measures of concordances such as the Kendall's Tau, Gini's Mean Difference,*

Spearman's Rho, and a weak interpretation of the Weak Concordance are used to identify these generic trends. The concept is validated using Financial Time Series data (S&P 500 Index) as the sample data set.

36. Kumara MPTR, Fernando WMS, Perera JMCU, Philips CHC. Time Series Prediction Algorithms: Literature Review (2013), Sri Lanka.

37. Martínez-Álvarez F, Troncoso A, Asencio-Cortés G, Riquelme J. A Survey on Data Mining Techniques Applied to Electricity-Related Time Series Forecasting. *Energies.* 2015; 8(12):13162-13193.

--- *Application of data mining techniques to time series forecasting. Provide a compact mathematical formulation of the mainly used techniques. Review the latest works of time series forecasting and, as case study, those related to electricity price and demand markets.*

38. McShane BB. *Machine Learning Methods with Time Series Dependence* [PhD Thesis]: Department of Managerial Science and Applied Economics, University of Pennsylvania; 2010.

--- *PrAGMaTiSt: Prediction and Analysis for Generalized Markov Time Series of States, a methodology which enhances classification algorithms so that they can accommodate sequential data. The PrAGMaTiSt can model a wide variety of time series structures including arbitrary order Markov chains, generalized and transition dependent generalized Markov chains, and variable length Markov chains.*

39. Mitchell S. *The Application of Machine Learning Techniques to Time-Seires Data* [Master's Thesis]: Department of Computing and Mathematical Sciences, University of Waikato; 1995.

--- *New methods for discovering knowledge in real world time-series data. Two complementary approaches were investigated: 1) manipulation of the original dataset into a form that is usable by conventional similarity-based learners; and 2) using sequence identification techniques to learn the concepts embedded in the database.*

40. Pritzsche U. *Benchmarking of Classical and Machine-Learning Algorithms (with special emphasis on Bagging and Boosting Approaches) for Time Series Forecasting* [Master's Thesis]: Deoartment of Statistics, Ludwig Maximilians Universitat Munchen; 2015.

--- *Evaluates the time series forecast capability of several Machine Learning approaches: Neural Nets, Random Forests, Kernel Machines (Support Vector Machines and Gaussian Processes), tree-based and component-wise linear and spline-based Boosting, by a comparison with classical ARIMA and ETS models.*

41. Stepnicka M, Cortez P, Donate JP, Stepnickova L. Forecasting seasonal time series with computational intelligence: on recent methods and the potential of their combinations. *Expert Systems with Applications.* 2013; 40(6):1981-1992.

--- *Novel methods for multi-step seasonal time series forecasting. All the presented methods stem from computational intelligence techniques: evolutionary artificial neural networks, support vector machines and genuine linguistic fuzzy rules.*

42. Zhang G, Eddy Patuwo B, Hu MY. Forecasting with artificial neural networks. *International Journal of Forecasting.* 1998; 14(1):35-62.

--- *Survey of state-of-the-art of ANN applications in forecasting.*

Extension Rules

43. Ben Taieb S. Strategies for multi-step time series forecasting. Paper presented at: International Symposium on Forecasting (2011); Prague.

44. Ben Taieb S. *Machine learning strategies for multi-step-ahead-time series forecasting* [PhD Thesis]: Departement d'Informatique, Universite Libre de Bruxelles; 2014.

45. Ben Taieb S, Bontempi G, Atiya AF, Sorjamaa A. A review and comparison of strategies for multi-step ahead time series forecasting based on the NN5 forecasting competition. *Expert Systems with Applications.* 2012; 39(8):7067-7083.

46. Ben Taieb S, Hyndman RJ. Recursive and direct multi-step forecasting: the best of both worlds: a Working paper. (2012), Monash University, Department of Econometrics and Business Statistics.

47. Bontempi G. Machine Learning Strategies for Time Series Prediction. Paper presented at: European Business Intelligence Summer School (2013); Brussels.

48. Bontempi G, Ben Taieb S, Le Borgne Y-A. Machine Learning Strategies for Time Series Forecasting. *Lecture Notes in Business Information Processing.* Vol 138: Springer Berlin Heidelberg; 2013:62-77.

--- *Prepare time series data for machine learning. Formalization of time series to cross sectional representation. Recursive, direct, and other forecasting strategies.*

Training & Testing Rules

49. Bergmeir C, Benitez JM. Forecaster Performance Evaluation with Cross-validation and Variants. Paper presented at: International Conference on Intelligent Systems Design and Applications (2011).

50. Bergmeir C, Benítez JM. On the use of cross-validation for time series predictor evaluation. *Information Sciences.* 2012; 191:192-213.

51. Bergmeir C, Costantini M, Benítez JM. On the usefulness of cross-validation for directional forecast evaluation. *Computational Statistics & Data Analysis.* 2014; 76:132-143.

52. Borra S, Di Ciaccio A. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. *Computational Statistics & Data Analysis.* 2010; 54(12):2976-2989.

53. Burman P, Chow E, Nolan D. A cross-validatory method for dependent data. *Biometrika.* 1994; 81(2):351-358.

54. Hyndman RJ. Measuring forecast accuracy Mar 31, 2014 (2014).

Metrics

55. Ahlburg DA, Chatfield C, Taylor SJ, et al. A Commentary on Error Measures. *International Journal of Forecasting.* 1992; 8(1):99-111.

--- *Comments on Armstrong & Collopy, Fildes choice of appropriate metrics. Survey of 17 papers dealing with population forecasts, they used of MAPE, RMSE, RMSPE, Theil's U - none justified the choice of error measure.*

56. Armstrong JS. Evaluating Forecasting Methods. In: Armstrong JS, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*: Kluwer Academic Publishers; 2001.

--- *Checklist of 32 principles. Evaluation by four tests: testing assumptions, testing data and methods, replicating outputs, assessing outputs. Principles based on commonly accepted methodologucal procedures. Do not use R2, do not use MSE, do not use within-sample fit.*

57. Armstrong JS. Selecting Forecasting Methods. In: Armstrong JS, ed. *Principles of Forecasting: A Handbook for Researchers and Practitioners*: Kluwer Academic Publishers; 2001.

--- *Flowchart to guide among ten forecasting methods. Six ways to select forecasting methods: convenience, market popularity, structured judgment, statistical criteria, relative track records, guidelines from prior research.*

58. Armstrong JS, Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting.* 1992; (8):69-80.

--- *Evaluates measures for making comparisons of errors across time series. Judged on reliability, construct validity, sensitivity to small changes, protection against outliers, relationship to decision making. Recommend GMRAE, MdRAE, MdAPE. Not RMSE.*

59. Armstrong JS, Fildes R. On the Selection of Error Measures for Comparisons Among Forecasting Methods. *Journal of Forecasting.* 14:67-71.

--- *Criticsm of Generalized Forecast Error Second Moment (GFESM) as an improvement to the Mean Square Error in comparing forecasting performance across data series.*

60. Davydenko A. How to Measure the Quality of Demand Forecasts Efficiently: a New Class of Forecasting Performance Metrics. Paper presented at: LUMS Inaugural Research Conference; Jun 30, 2010 (2010); Lancaster, United Kingdom.

61. Hyndman RJ. Another Look at Forecast Accuracy Metrics for Intermittent Demand. *Foresight: The International Journal of Applied Forecasting.* 2006; (4):43-46.

--- *Summariy of forecast accuracy metrics and their potential failings. Introduces the mean absolute scaled error (MASE).*

62. Hyndman RJ. Errors on percentage errors. *Hyndsight: A blog by Rob J Hyndman.* (Apr 16, 2014). http://robjhyndman.com/hyndsight/smape/.

--- *Encourages use of MAPE and MASE.*

63. Hyndman RJ, Koehler AB. Another look at measures of forecast accuracy. *International Journal of Forecasting.* 2006; 22(4):679-688.

--- *Comparison of measures of accuracy of univariate time series forecasts. Methods used in the M-competition and the M3-competition, and many of the measures recommended by previous authors on this topic, are found to be degenerate in commonly occurring situations.*

64. Kim S, Kim H. A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting.* 2016; 32(3):669-679.

--- *Proposal for a new measure of forecast accuracy called the mean arctangent absolute percentage error (MAAPE). MAPE has the significant disadvantage in that it produces infinite or undefined values for zero or close-to-zero actual values.*

65. Shcherbakov MV, Brebels A, Shcherbakov NL, Tyukov AP, Janovsky TA, Kamaev VAe. A Survey of Forecast Error Measures. *World Applied Sciences Journal.* 24:171-176.

--- *Review of the commonly used forecast error measurements. 7 groups: absolute forecasting errors, measures based on percentage errors, symmetric errors, measures based on relative errors, scaled errors, relative measures and other error measures.*

66. Syntetos AA, Boylan JE. The accuracy of intermittent demand estimates. *International Journal of Forecasting.* 2005; 21(2):303-314.

--- *Four forecasting methods, Simple Moving Average (SMA, 13 periods), Single Exponential Smoothing (SES), Croston's method, and a new method (based on Croston's approach) recently developed by the authors, are compared on 3000 real intermittent demand data series from the automotive industry. The mean signed and relative geometric root-mean-square errors are shown to meet the theoretical and practical requirements of intermittent demand.*

Tools

67. Chang C-C, Lin C-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology.* 2011; 2(3):1-27.

68. Cortez P. *A tutorial on using the rminer R package for data mining tasks.* Department of Information Systems, ALGORITMI Research Centre, Engineering School, University of Minho; (2015). http://hdl.handle.net/1822/36210.

69. Cortez P. *rminer: Data Mining Classification and Regression Methods.* (2015). http://CRAN.R-project.org/package=rminer.

70. Gromping U. Relative Importance for Linerar Regression in R: The Package relaimpo. *Journal of Statistical Software.* 2006; 17(1):1-27.

71. Machine Learning Group at the University of Waikato. Weka 3: Data Mining Software in Java. 2016. http://www.cs.waikato.ac.nz/ml/weka/.

72. R Core Team. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing; (2014). http://www.R-project.org.

73. Wickham H. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York; 2009.

Comparison Studies

74. Ahmed NK, Atiya AF, Gayar NE, El-Shishiny H. An Empirical Comparison of Machine Learning Models for Time Series Forecasting. *Econometric Reviews.* 2010; 29(5-6):594-621.

75. Wang W-C, Chau K-W, Cheng C-T, Qiu L. A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series. *Journal of Hydrology.* 2009; 374(3-4):294-306.

Reviews of Comparison Studies Focused on Electricity Demand

76. Hahn H, Meyer-Nieberg S, Pickl S. Electric load forecasting methods: Tools for decision making. *European Journal of Operational Research.* 2009; 199(3):902-907.

77. Hippert HS, Pedreira CE, Souza RC. Neural networks for short-term load forecasting: a review and evaluation. *IEEE Transactions on Power Systems.* 2001; 16(1):44-55.

78. Rui Y, El-Keib AA. A review of ANN-based short-term load forecasting models. Paper presented at: 27th Southeastern Symposium on System Theory; Mar 12-14, 1995 (1995).

79. Singh AK, Ibraheem, Khatoon S, Muazzam M, Chaturvedi DK. Load Forecasting Techniques and Methodologies: A Review. *2012 2nd International Conference on Power, Control and Embedded Systems (Icpces 2012).* 2012.

Comparison Studies Focused on Electricity Demand

80. Ahmad AS, Hassan MY, Abdullah MP, et al. A review on applications of ANN and SVM for building electrical energy consumption forecasting. *Renewable and Sustainable Energy Reviews.* 2014; 33:102-109.

--- *Review of building electrical energy forecasting using artificial intelligence methods such as support vector machine (SVM) and artificial neural network (ANN). Also, hybridization of the two forecasting methods, Group Method of Data Handling (GMDH), and Least Square Support Vector Machine (LSSVM aka GLSSVM) to forecast building electrical energy consumption.*

81. Aung Z, Toukhy M, Williams J, Sanchez A, Herrero S. Towards Accurate Electricity Load Forecasting in Smart Grids. Paper presented at: International Conferecne on Advances in Databases, Knowledge, and Data Applications (2012).

--- *Forecasting of future grid load (electricity usage) is an important task to provide intelligence to the smart gird. Proposal of a new data mining scheme to forecast the peak load of a particular consumer entity in the smart grid for a future time unit. Utilizes least-squares version of support vector regression with online learning strategy in our approach.*

82. Baliyan A, Gaurav K, Mishra SK. A Review of Short Term Load Forecasting using Artificial Neural Network Models. *Procedia Computer Science.* 2015; 48:121-125.

--- *Review of recently published research on different variants of artificial neural network in the field of short term load forecasting, especially the hybrid networks, which is a combination of neural network with stochastic learning techniques such as genetic algorithm(GA), particle swarm optimization (PSO) etc.*

83. Ben Taieb S. Forecasting Uncertainty in Electricity Smart Meter Data by Boosting Additive Quantile Regression. *IEEE Transactions on Smart Grid.* Submitted 2015.

--- *Proposal for a probabilistic forecasting method where a different quantile regression model is estimated for each quantile of the future distribution. Each model is estimated by boosting additive quantile regression. Compare approach with three benchmark methods on both aggregated and disaggregated scales using a smart meter dataset collected from 3639 households in Ireland at 30-minute intervals over a period of 1.5 years.*

84. Ben Taieb S, Huser R, Hyndman RJ, Genton MG. Probabilistic time series forecasting with boosted additive models: an application to smart meter data. SIGKDD Workshop on Mining and Learning from Time Series; 2015; Sydney, Australia.

--- *Quantile regression to forecast household. Ireland CER smart meter dataset, 250 meters (out of 3639), 30 min resolution, 18 months, not include heat/cool.*

85. Black JD. Wind Integration In New England. Paper presented at: 8th Annual Conference on the Electricity Industry; Mar 13, 2012 (2012).

86. Black JD. *Load Hindcasting: A Retrospective Regional Load Prediction Method Using Reanalysis Weather Data* [Master]: Mechanical Engineering, University of Massachusetts Amherst.

--- *The capacity value (CV) of a power generation unit indicates the extent to which it contributes to the generation system adequacy of a region's bulk power system. Given the capricious nature of the wind resource, determining wind generation's CV is nontrivial, but can be understood simply as how well its power output temporally correlates with a region's electricity load during times of system need. Both wind generation and load are governed by weather phenomena that exhibit variability across all timescales, including low frequency weather cycles that span decades. Thus, a data-driven determination of wind's CV should involve the use of long-term (i.e., multiple decades) coincident load and wind data.*

87. Borges CE, Penya YK, Fernandez I. Evaluating Combined Load Forecasting in Large Power Systems and Smart Grids. *IEEE Transactions on Industrial Informatics.* 2013; 9(3):1570-1577.

--- *Combined aggregative short-term load forecasting method for smart grids, a novel methodology that allows us to obtain a global prognosis by summing up the forecasts on the compounding individual loads.*

Detail three new approaches, namely bottom-up aggregation (with and without bias correction), top-down aggregation (with and without bias correction), and regressive aggregation. Experiment to compare the results, evaluating them with two datasets of real data and showing the feasibility of aggregative forecast combinations for smart grids.

88. Ceperic E, Ceperic V, Baric A. A Strategy for Short-Term Load Forecasting by Support Vector Regression Machines. *IEEE Transactions on Power Systems.* 2013; 28(4):4356-4364.

--- *Generic strategy for short-term load forecasting (STLF) based on the support vector regression machines (SVR). Two important improvements to the SVR based load forecasting method are introduced, i.e., procedure for generation of model inputs and subsequent model input selection using feature selection algorithms. Feature selection algorithms for automatic model input selection and the use of the particle swarm global optimization based technique for the optimization of SVR hyper-parameters reduces the operator interaction. To confirm the effectiveness of the proposed modeling strategy, the model has been trained and tested on two publicly available and well-known load forecasting data sets.*

89. Chen Y, Hong W-C, Shen W, Huang N. Electric Load Forecasting Based on a Least Squares Support Vector Machine with Fuzzy Time Series and Global Harmony Search Algorithm. *Energies.* 2016; 9(2):70.

--- *Proposal for a new electric load forecasting model by hybridizing the fuzzy time series (FTS) and global harmony search algorithm (GHSA) with least squares support vector machines (LSSVM), namely GHSA-FTS-LSSVM model. Fuzzy c-means clustering (FCS) algorithm is used to calculate the clustering center of each cluster. LSSVM is applied to model the resultant series, which is optimized by GHSA. Model is verified using experimental datasets from the Guangdong Province Industrial Development Database. Results are compared against autoregressive integrated moving average (ARIMA) model and other algorithms hybridized with LSSVM including genetic algorithm (GA), particle swarm optimization (PSO), harmony search, and so on.*

90. Cui C, Wu T, Hu M, Weir JD, Li X. Short-term building energy model recommendation system: A meta-learning approach. *Applied Energy.* 2016; 172:251-263.

--- *A generalized system framework which can recommend appropriate models to forecast the building energy profiles based on building characteristics. Proposal for a meta-learning based framework, termed Building Energy Model Recommendation System (BEMR). Based on the building's physical features and statistical and time series meta-features extracted from the operational data and energy consumption data. Three sets of experiments on 48 test buildings and one real building were conducted.*

91. De Felice M, Xin Y. Short-Term Load Forecasting with Neural Network Ensembles: A Comparative Study [Application Notes]. *IEEE Computational Intelligence Magazine.* 2011; 6(3):47-56.

92. De Silva D, Yu X, Alahakoon D, Holmes G. A Data Mining Framework for Electricity Consumption Analysis From Meter Data. *IEEE Transactions on Industrial Informatics.* 2011; 7(3):399-407.

--- *Novel data mining framework for the exploration and extraction of actionable knowledge from data generated by electricity meters. Framework incorporates functionality for interim summarization and incremental analysis using intelligent techniques. The proposed Incremental Summarization and Pattern Characterization (ISPC) framework demonstrates this capability. Stream data is structured in a data warehouse based on key dimensions enabling rapid interim summarization. Independently, the IPCL algorithm incrementally characterizes patterns in stream data and correlates these across time. Eventually, characterized patterns are consolidated with interim summarization to facilitate an overall analysis and prediction of energy consumption trends. Results of experiments conducted using the actual data from electricity meters confirm applicability of the ISPC framework.*

93. Ding N, Benoit C, Foggia G, Besanger Y, Wurtz F. Neural Network-Based Model Design for Short-Term Load Forecast in Distribution Systems. *IEEE Transactions on Power Systems.* 2016; 31(1):72-81.

--- *Neural network beats time series model on French disribution system data.*

94. Ertugrul ÖF. Forecasting electricity load by a novel recurrent extreme learning machines approach. *International Journal of Electrical Power & Energy Systems.* 2016; 78:429-435.

--- *Proposal fopr recurrent extreme learning machine (RELM) to forecast electricity load more accurately. In RELM, extreme learning machine (ELM), which is a training method for single hidden layer feed forward neural network, was adapted to train a single hidden layer Jordan recurrent neural network. Electricity Load Diagrams 2011–2014 dataset was employed to evaluate and validate the proposed approach.*

95. Guo Y-C. Knowledge-Enabled Short-Term Load Forecasting Based on Pattern-Base Using Classification & Regression Tree and Support Vector Regression. 2009:425-429.

--- *A new model of short-term load forecasting based on pattern-base. It recognizes the different patterns of daily load according such features as weather and date type by means of data support vector mining technology of classification and regression tree. It sets up pattern-bases which copmprise daily load data sequence with highly similar establishes features. It uses a regression forecasting model based on the pattern-base which matches to the forecasting day.*

96. Gvaladze S. *Evaluating methods for time-series forecasting: Applied to energy consumption predictions for Hvaler* [Master's Thesis]: Department of Computer Science, Ostfold University College; 2015.

--- *Comparison of different techniques for electricity consumption data of the one station in Hvaler. Description of the several approaches for similar problems, general description of statistical and machine learning models and applying those models for specific time-series data. The methods can be grouped as statistical and machine learning. Discussion of main challenges when dealing with time-series data. The following models are applied to electricity consumption data: auto-regressive integrated moving average (ARIMA), linear regression, decision and model trees, regression trees, support vector machines(SVM), k-nearest neighbor, neural network.*

97. Hsiao Y-H. Household Electricity Demand Forecast Based on Context Information and User Daily Schedule Analysis From Meter Data. *IEEE Transactions on Industrial Informatics.* 2015; 11(1):33-43.

--- *A novel approach to model the very short-term load of individual households based on context information and daily schedule pattern analysis. Several daily behavior pattern types were obtained by analyzing the time series of daily electricity consumption, and context features from various sources were collected and used to establish a rule set for use in anticipating the likely behavior pattern type of a*

specific day. Electricity consumption volume prediction model was developed for each behavior pattern type to predict the load at a specific time point in a day. This study was concerned with solving the VSTLF for individual households in Taiwan.

98. Hu Z, Bao Y, Chiong R, Xiong T. Mid-term interval load forecasting using multi-output support vector regression with a memetic algorithm for feature selection. *Energy.* 2015; 84:419-431.

--- *Models and forecasts mid-term interval loads up to one month in the form of interval-valued series consisting of both peak and valley points by using MSVR (Multi-output Support Vector Regression). Also, an MA (Memetic Algorithm) based on the firefly algorithm is used to select proper input features among the feature candidates, which include time lagged loads as well as temperatures. The capability of this proposed interval load modeling and forecasting framework to predict daily interval electricity demands is tested through simulation experiments using real-world data from North America and Australia.*

99. Humeau S, Wijaya TK, Vasirani M, Aberer K. Electricity load forecasting for residential customers: Exploiting aggregation and correlation between households. Paper presented at: Sustainable Internet and ICT for Sustainability (SustainIT) (2013); Palermo, Italy.

--- *Predicting electricity consumption at very low scales. Addresses the problem of improving consumption forecasting by using the statistical relations between consumption series. This is done both at the household and district scales (hundreds of houses), using various machine learning techniques, such as support vector machine for regression (SVR) and multilayer perceptron (MLP). First, determine which algorithm is best adapted to each scale, then, try to find leaders among the time series, to help short-term forecasting. Improve the forecasting for district consumption by clustering houses according to their consumption profiles.*

100. Hyndman RJ, Fan S. Density forecasting for long-term peak electricty demand: a manuscript. Aug 7, 2008 (2008), Monash University.

--- *Proposal for a new methodology to forecast the density of long-term peak electricity demand. Peak electricity demand in a given season is subject to a range of uncertainties, including underlying population growth, changing technology, economic conditions, prevailing weather conditions (and the timing of those*

conditions), as well as the general randomness inherent in individual usage. It is also subject to some known calendar effects due to the time of day, day of week, time of year, and public holidays. Uses semi-parametric additive models to estimate the relationships between demand and the driver variables, including temperatures, calendar effects and some demographic and economic variables, forecasts the demand distributions using a mixture of temperature simulation, assumed future economic scenarios, and residual bootstrapping. Compares the forecast results with the actual demand of the summer 2007/08.

101. Jain BM, Nigam MK, Tawari PC. Curve fitting and regression line method based seasonal short term load forecasting. Paper presented at: World Congress on Information and Communication Technologies (WICT); Oct 30, 2012 (2012).

102. Jurado S, Nebot À, Mugica F, Avellana N. Hybrid methodologies for electricity load forecasting: Entropy-based feature selection with machine learning and soft computing techniques. *Energy.* 2015; 86:276-291.

--- *Compares the accuracy of different machine learning methodologies for the hourly energy forecasting in buildings. Proposal for a hybrid methodology that combines feature selection based on entropies with soft computing and machine learning approaches, i.e. Fuzzy Inductive Reasoning, Random Forest and Neural Networks. Compare with a traditional statistical technique ARIMA (AutoRegressive Integrated Moving Average). Generates fast and reliable models, with low computational costs that could be embedded in second generation of smart meters, where they could generate on-site electricity forecasting of the next hours, or trade the excess of energy.*

103. Jurado S, Peralta J, Nebot A, Mugica F, Cortez P. Short-term electric load forecasting using computational intelligence methods. Paper presented at: International Conference on Fuzzy Systems (FUZZ) (2013); Hyderabad.

--- *Introduces several methods for short-term electric load forecasting. All the presented methods stem from computational intelligence techniques: Random Forest, Nonlinear Autoregressive Neural Networks, Evolutionary Support Vector Machines and Fuzzy Inductive Reasoning.*

104. Kandil N, Wamkeue R, Saad M, Georges S. An efficient approach for short term load forecasting using artificial neural networks. *International Journal of Electrical Power & Energy Systems.* 2006; 28(8):525-530.

--- *Artificial neural networks (ANN) for short term load forecasting using real load and weather data from the Hydro-Quebec databases, where three types of variables were used as inputs to the neural network: (a) hour and day indicators, (b) weather related inputs and (c) historical loads. Demonstrate ANN capabilities in load forecasting without the use of load history as an input.*

105. Koo B-g, Lee S-W, Kim W, Park JH. Comparative Study of Short-Term Electric Load Forecasting. 2014:463-467.

--- *Short-term electric load forecasting using three methods: ANN, SES, GMDH. We carried out 1-day ahead prediction for two weeks, January 10 to 16, March 14 to 20, 2011 using hourly Korean electric load data.*

106. Lang K, Zhang M, Yuan Y. Improved Neural Networks with Random Weights for Short-Term Load Forecasting. *PLoS One.* 2015; 10(12).

--- *Proposal for a new forecasting model based on the improved neural networks with random weights (INNRW). The key is to introduce a weighting technique to the inputs of the model and use a novel neural network to forecast the daily maximum load. Eight factors are selected as the inputs. A mutual information weighting algorithm is then used to allocate different weights to the inputs. The neural networks with random weights and kernels (KNNRW) is applied to approximate the nonlinear function between the selected inputs and the dailymaximum load due to the fast learning speed and good generalization performance. In the application of the daily load in Dalian, the result of the proposed INNRW is compared with several previously developed forecasting models.*

107. Mehmood ST, El-Hawary M. Performance Evaluation of New and Advanced Neural Networks for Short Term Load Forecasting. 2014:202-207.

--- *New and advanced neural network (NN) architectures to perform STLF. Two hybrid and two 3-layered NN architectures are introduced. Each network is individually tested to generate weekday and weekend forecasts using data of Nova Scotia, Canada.*

108. Mirowski P, Chen S, Kam Ho T, Yu C-N. Demand Forecasting in Smart Grids. *Bell Labs Technical Journal.* 2014; 18(4):135-158.

--- *Exploit rich, multi-year, and high-frequency annotated data collected via a metering infrastructure to perform STLF on aggregates of power meters in a mid-sized city. For smart meter aggregates complemented with geo-specifi c weather data, benchmark several state-of-the-art forecasting algorithms, including kernel methods for nonlinear regression, seasonal and temperature-adjusted auto-regressive models, exponential smoothing and state-space models.*

109. Rahman SM. *Data Driven Models Applied in Building Load Forecasting for Residential and Commercial Buildings* [Master's Thesis]: Department of Mechanical Engineering, University of Texas; 2015.

--- *Three different state-of-art machine learning methods, i.e., Artificial Neural Network, Support Vector Regression and Gaussian Process Regression are applied in hour ahead and 24–hour ahead building energy forecasting. The work uses four residential buildings and one commercial building located in Downtown, San Antonio as test-bed using energy consumption data from those buildings monitored in real-time. Uncertainty quantification analysis is conducted to understand the confidence in each forecast using Bayesian Network. Using a combination of weather variables and historical load, forecasting is done in a supervised way based on a moving window training algorithm.*

110. Rahman SM, Dong B, Vega R. Machine Learning Approach Applied in Electricity Load Forecasting: Within Residential Houses Context. *ASHRAE Transactions.* 2015; 121.

--- *Explore the effectiveness of four current, state-of-art machine learning methods applied in hour and day ahead (24 hours) electricity load forecasting in a residential context: traditional Artificial Neural Network (ANN), Support Vector Regression (SVR), Gaussian Process Regression (GPR) and Bayesian Network (BN). The paper uses four houses located in San Antonio as a test-bed using energy consumption data from those houses monitored in real-time. Uncertainty quantification analysis is conducted on the forecasts to understand the error bands. Using a combination of weather variables and historical load, forecasting is done in a supervised way based on a moving window training algorithm.*

111. Soares LJ, Medeiros MC. Modeling and forecasting short-term electricity load: A comparison of methods with an application to Brazilian data. *International Journal of Forecasting.* 2008; 24(4):630-644.

--- *Forecasting model for the hourly electricity load in the area covered by an electric utility located in the southeast of Brazil.*

112. Son H, Kim C. Forecasting Short-term Electricity Demand in Residential Sector Based on Support Vector Regression and Fuzzy-rough Feature Selection with Particle Swarm Optimization. *Procedia Engineering.* 2015; 118:1162-1168.

--- *Model for one-month-ahead forecast of electricity demand in residential sector. 20 influential variables are taken into account, including monthly electricity consumption, 14 weather variables, and 5 social variables. Based on support vector regression and fuzzy-rough feature selection with particle swarm optimization algorithms, the proposed method established a model with variables that relate to the forecast without ignoring some of these variables one may inevitably lead to forecasting errors. The proposed forecasting model was validated using historical data from South Korea. Its time period was from January 1991 to December 2012. The first 240 months were used for training and the remaining 24 for testing. The performance was evaluated using MAPE, MAE, RMSE, MBE, and UPA values.*

113. Taylor JW, de Menezes LM, McSharry PE. A comparison of univariate methods for forecasting electricity demand up to a day ahead. *International Journal of Forecasting.* 2006; 22(1):1-16.

--- *Compares the accuracy of six univariate methods for short-term electricity demand forecasting for lead times up to a day ahead. The methods considered include the recently proposed exponential smoothing method for double seasonality and a new method based on principal component analysis (PCA). The methods are compared using a time series of hourly demand for Rio de Janeiro and a series of half-hourly demand for England and Wales.*

114. Taylor JW, McSharry PE. Short-Term Load Forecasting Methods: An Evaluation Based on European Data. *IEEE Transactions on Power Systems.* 2007; 22(4):2213-2219.

--- *Uses intraday electricity demand data from 10 European countries as the basis of an empirical comparison of univariate methods for prediction up to a day-ahead. Forecasting methods considered in the study include: ARIMA modeling; periodic AR modeling; an extension for double seasonality of Holt-Winters exponential alternative exponential smoothing formulation; and a method based on the principal component analysis (PCA) of the daily demand profiles.*

115. Tsekouras GJ, Kanellos FD, Mastorakis N. Short Term Load Forecasting in Electric Power Systems with Artificial Neural Networks. In: Mastorakis N, ed. *Computational Problems in Science and Engineering.* Vol 343: Springer International Publishing Switzerland; 2015:19-58.

--- *Introduction to Artificial Neural Networks and their usage in forecasting the load demand of electric power systems. Several of the major training techniques are described with their pros and cons being discussed. Feed-forward ANNs are used for the short-term forecasting of the Greek Power System load demand. Various ANNs with different inputs, outputs, numbers of hidden neurons are examined, and techniques for their optimization are proposed.*

116. Wang P, Liu B, Hong T. Electric load forecasting with recency effect: a big data approach. *International Journal of Forecasting.* 2016; 32(3):585-597.

--- *"Recency effect", a term originated from psychology, denotes that electricity demand is affected by the temperatures of preceding hours. Comprehensive study on modeling recency effect through a big data approach. How many lagged hourly temperatures and/or moving average temperatures are needed in a regression model to fully capture recency effect without compromising the forecasting accuracy? Uses the case study based on data from the load forecasting track of the Global Energy Forecasting Competition 2012.*

117. Wijaya TK, Vasirani M, Humeau S, Aberer K. Cluster-based aggregate forecasting for residential electricity demand using smart meter data. Paper presented at: IEEE International Conference on Big Data; Oct. 29 2015-Nov. 1 2015 (2015); Santa Clara, CA.

--- *Focuses on short-term (1 and 24 hour ahead) electricity demand forecasting for residential customers at the individual and aggregate level. Build a feature universe, and then apply Correlation-based Feature Selection to select features relevant to each household. Smart meter data can be used to obtain aggregate forecasts with higher accuracy using the so-called Cluster-based Aggregate Forecasting (CBAF) strategy, i.e., by first clustering the households, forecasting the clusters' energy consumption separately, and finally aggregating the forecasts.*

118. Yang Y, Meng Y, Xia Y, Lu Y, Yu H. An Efficient Approach for Short Term Load Forecasting. Paper presented at: International Multiconferecne of Engineers and Computer Scientists; Mar 16-18, 2011; Hong Kong.

--- *Proposeal for a novel short term load forecasting approach based on training data selection. The load curve of a time interval before the target hour is regard as the benchmark of training data instead of the cluster center of all historical data used in previous studies. The load curves are normalized for comparison to the benchmark.*

Comaprison Studies Focused on Electricity Demand & Data Strategies

119. Black JD, Henson WLW. Hierarchical Load Hindcasting Using Reanalysis Weather. *IEEE Transactions on Smart Grid.* 2014; 5(1):447-455.

120. Fan S, Wu Y-K, Lee W-J. Comparative study on load forecasting technologies for different geographical distributed loads. Paper presented at: Power and Energy Society General Meeting; July 24-29, 2011 (2011).

121. Grolinger K, L'Heureux A, Capretz MAM, Seewald L. Energy Forecasting for Event Venues: Big Data and Prediction Accuracy. *Energy and Buildings.* 2016; 112:222-233.

122. Hayes B, Gruber J, Prodanovic M. Short-Term Load Forecasting at the local level using smart meter data. 2015:1-6.

123. Sevlian R, Rajagopal R. Short Term Electricity Load Forecasting on Varying Levels of Aggregation. *Submitted to Transactions on Power Systems.* 2014.

124. Sevlian RA, Rajagopal R. A model for the effect of aggregation on short term load forecasting. Paper presented at: 2014 IEEE PES General Meeting, Conference, and Exposition; Jul 27-31, 2014 (2014).

Industry Reports

125. Aguirre V. *2016 Preliminary Integrated Resource Plan.* Tucson Electric Power; Mar 1, 2016 (2016).

126. Dermot N, Blaney G. *Regulator's 2012 National Report to the European Commission.* Commission for Energy Regualtion (CER) Ireland; Jul 2012 (2012).

127. Erhlich D, Winkler E, Black JD. *Draft 2016 CELT ISO-NE Annual Energy and Summer Peak Forecast.* ISO New England; Mar 22, 2016 (2016).

128. ESB Networks. *Electricity Smart Metering Technology Trials Findings Report.* Commission for Energy Regulation; May 16, 2011 (2011).

129. Itron I. *2014 Forecasting Benchmark Survey.* Itron; Sep 16, 2014 (2014).

130. Sheehan M. *2014 Integrated Resource Plan.* Tucson Electric Power; Apr 1, 2014 (2014).

Electricity Pricing

131. Avalon Energy Services. Day-Ahead and Real-Time Pricing During a Heat Wave. *Avalon Energy Services.* 2013 (July 22, 2013). http://www.avalonenergy.us/blog/?p=691.

132. Giordano V, Onyeji I, Fulli G, Jimenez MS, Filliou C. *Guidelines for Cost Benefit Analysis of Smart Metering Deployment.* Eurpean Commission Joint Research Center, Institute for Energy and Transport; (2012). 978-92-79-22323-5. http://www.jrc.ec.europa.eu.

133. Hamachi Lacommare K, Eto JH. Cost of Power Interruptions to Electrcity Consumers in the United States. *Energy.* 2006; 31(12):1845-1855.

Forecasting Competitions

134. Athanasopoulos G, Hyndman RJ. The value of feedback in forecasting competitions. *International Journal of Forecasting.* 2011; 27(3):845-849.

135. Clements MP, Hendry DF. Explaining the results of the M3 forecasting competition. *International Journal of Forecasting.* 2001; 17(4):550-554.

136. Crone SF, Hibon M, Nikolopoulos K. Advances in forecasting with neural networks? Empirical evidence from the NN3 competition on time series prediction. *International Journal of Forecasting.* 2011; 27(3):635-660.

137. IEEE PES Announces the Eight Winning Teams for the Global Energy Forecasting Competition 2012 [press release]. IEEE Power & Energy Society, Sep 30, 2013.

138. IEEE Power & Energy Society Launches the Global Energy Forecasting Competition 2014 [press release]. IEEE Power & Energy Society, Aug 15, 2014 2014.

139. Hong T, Pinson P, Fan S. Global Energy Forecasting Competition 2012. *International Journal of Forecasting.* 2014; 30(2):357-363.

140. IEEE Power & Energy Society Announces the Winning Universities and Teams for the Global Energy Forecasting Competition 2014 [press release]. IEEE Power & Energy Society.

141. Makridakis S, Hibon M. The M3-Competition: results, conclusions and implications. *International Journal of Forecasting.* 2000; 16:451-476.

142. Nedellec R, Cugliari J, Goude Y. GEFCom2012: Electric load forecasting and backcasting with semi-parametric models. *International Journal of Forecasting.* 2014; 30(2):375-381.

143. Xie J, Hong T. GEFCom2014 probabilistic electric load forecasting: An integrated solution with forecast combination and residual simulation. *International Journal of Forecasting.* 2016; 32(3):1012-1016.

Relative Importance

144. Bi J. A Review of Statistical Methods for Determination of Relative Importance of Correlated Predictors and Identification of Drivers of Consumer Liking. *Journal of Sensory Studies.* 2012; 27(2):87-101.

145. Grömping U. Estimators of Relative Importance in Linear Regression Based on Variance Decomposition. *The American Statistician.* 2007; 61(2):139-147.

146. Grömping U. Variable importance in regression models. *Wiley Interdisciplinary Reviews: Computational Statistics.* 2015; 7(2):137-152.

147. Lindeman RH, Merenda PF, Gold RZ. *Introduction to bivariate and multivariate analysis.* Scott, Foresman; 1980.

Entropy & Forecastability

148. Catt P. Entropy as an A Priori Indicator of Forecastability Nov 2014 (2014).

149. Maasoumi E, Racine J. Entropy and Predictability of Stock Market Returns. *Journal of Econometrics.* 2002; 107:291-312.

150. Pincus SM. Approximate entropy as a measure of system complexity. *Proceedings of the National Academy of Sciences.* 1991; 88(6):2297-2301.

151. Wu S-D, Wu C-W, Lin S-G, Wang C-C, Lee K-Y. Time Series Analysis Using Composite Multiscale Entropy. *Entropy.* 2013; 15(3):1069-1084.

152. Yentes JM, Hunt N, Schmid KK, Kaipust JP, McGrath D, Stergiou N. The appropriate use of approximate entropy and sample entropy with short data sets. *Ann Biomed Eng.* 2013; 41(2):349-365.

Datasets

153. Australian Government. Smart-Grid Smart-City Customer Trial 2012-2014: Electricity use interval reading. 2012-14. Available from: https://data.gov.au/organization/department-of-industry.

154. Ben Taieb S. Data from the M4 Time Series Forecasting Competition. CRAN; 2016. Available from: https://github.com/bsouhaib/M4comp.

155. Commission for Energy Regulation (CER). CER Smart Metering Project | Electricity Customer Behaviour Trial, 2009-2010: Smart meter read data. Irish Social Science Data Archive (ISSDA). Available from: http://www.ucd.ie/issda/data/commissionforenergyregulationcer/.

156. Forecasters IIo. M3 Competition Dataset. International Institute of Forecasters; 2000. Available from: https://forecasters.org/resources/time-series-data/m3-competition/.

157. Pacific Gas and Electric. PG&E Energy Data Request Program. Available from: https://www.pge-energydatarequest.com/.

158. University of Texas at Austin. Pecan Street Demonstration. In: Austin UoTa, ed. Available from: https://dataport.pecanstreet.org/.

159. Weather Underground. Weather History for Dublin, Ireland. Available from: www.wunderground.com.