Horse and Cart
Methodology for Electricity Load and Price Forecast

Overview
The Homochronic Recessed Smoothing Ensemble (Horse) and Corrective Autoregression Transfer function (Cart) methodology provides a general framework for electricity load and price forecasts. Agnostic to input variables or Independent Systems Operators (ISO), the Horse model identifies the input variables most correlated to the dependent variable, filters the input data based on the time of year and time of day, and calculates a forecast based on previous observations. Systemic errors in the forecast are captured by the Cart.

Data Preparation
Prior to the initial Horse estimate, the input data is scaled and normalized for load, day-ahead price, and real-time price forecasts. If there is linear growth or decline in load due to population growth or industrial demand, the annual data is sorted by season and scaled by the results of a linear regression model. For day-ahead and real-time price, however, the past data is simply scaled by the central tendency of the current season and hour. Electricity prices are affected by gas, coal, nuclear, solar, and wind prices, all of which change from year to year. Subsequently, electricity prices change in a nonlinear manner. To capture these changes, the central tendency of a rolling window of current electricity prices at a given hour is calculated along with the central tendency at given hours of previous years. The electricity prices from the past are then scaled by the ratio of the current central tendency to the past central tendency. For example, if the median of real-time prices for January 2020 at hour 12 is $20 and it was $30 for hour 12 in January 2014, the real-time prices from January 2014 at hour 12 will be scaled by a factor of two thirds.

Horse
Electricity consumption, like most human activities, is ordered by time of day, day of the week, and week of the year. The Horse model leverages this basic insight to estimate a forecast based on the past observations scaled in the manner described above and then normalized. A rolling correlation of all the prepared input data is sorted relative to the dependent variable. For example, in the load forecast, the prepared input data may include the average temperature by hour of the weather stations within a given forecast zone, as well as heat index, cooling-degree days, and other derived data. The independent variables are identified as those most correlated to the dependent variable, the number of which is based on the input parameter known as correlated variable number. The prepared data is then filtered by the time of the year, day of the week (either weekday or weekend), and hour of the day.
The independent variables are scaled by the rolling correlation scores to capture their relative importance to the dependent variable. For example, if the rolling correlation of heat index to electricity load is 0.9, while the cooling-degree day correlation is 0.86, the filtered and normalized input data is scaled by those factors. The normalized independent variable forecast data is vectorized (for example, weather-forecast data in the load model) and the Euclidean distance to previous observations of all past filtered and normalized observations is calculated and sorted. If prior observations of independent variables $X_i = \{X_{i1}, X_{i2}, \ldots, X_{in}\}$ and their forecast values $X_f = \{X_{f1}, X_{f2}, \ldots, X_{fn}\}$ represent two points in Euclidean $n$-space, then distance from $X_i$ to $X_f$ is given by the Pythagorean formula:

$$d = \sqrt{(X_{i1} - X_{f1})^2 + (X_{i2} - X_{f2})^2 + \cdots + (X_{in} - X_{fn})^2}$$

The minimum distances are selected based on the sample-size input parameter. A sample of the scaled dependent variables on the dates associated with the minimum distances are retrieved. For example, if the independent variables for a forecast of hour 12 in July of the current year are identified as cooling-degree days, heat index, and air temperature, and the minimum Euclidean distance of the vectorized observations are from July 21, 2015, then the scaled load from July 21, 2015, hour 12 is the first observation in the sample. The Horse estimate at this hour is simply the mean of the observations in the sample space. If the sample size is three and the scaled prior observations of the load are 10,000, 11,000, and 12,000 MW, the Horse estimate is 11,000 MW. An example of the Horse estimate is shown in Exhibit 1.

**Exhibit 1** ERCOT Coastal Load Horse Forecast vs. Actual Load for July 5, 2019


There is also a momentum function in the Horse estimate to capture the momentum effects of prior and posterior hours on a given hour estimate. This is especially important in capturing price spikes. The same
logic of minimizing the distance to the forecast data is at work; the only difference being that the
distances of prior and posterior hours are factored into the overall minimum distance calculation.

**Cart**
The errors between the Horse forecast and actual observations are calculated and continuously updated. As shown in Exhibit 2, these errors are autocorrelated and contain information about newly emergent developments in the system being modeled. The Cart simply takes the time series of those errors and builds a seven-day forecast of errors from an autoregression model. These forecast errors are added into the Horse estimate. This is the Horse and Cart forecast. Both the Horse and the Horse and Cart forecasts are passed through a Savitzky-Golay filter to increase the forecast precision without distorting the overall signal of the forecast. The filtered forecasts are called the smoothed Horse forecast and the smoothed Horse and Cart forecast, respectively.

**Exhibit 2** ERCOT Coastal Load Horse Forecast Error for July 2019

Since there is typically no serial correlation in the price estimates, the Horse and Cart usually introduces more error in the day-ahead and real-time price forecasts. The Cart does typically reduce the error in load forecasts, however, as can be seen in Exhibit 3.

**Model Summary**
The following steps are taken in the Horse and Cart forecast model:

1. Input data is retrieved. For load models this includes historical weather data as well as derived data. For price models this includes historical load data, tie flow data, wind generation data. All models can include proprietary time-series data as well.
2. Load and price data are scaled through either linear regression or through the ratio of central tendencies.
3. All data except the dependent variable is z-scored.
4. A rolling correlation of the input variables to the dependent variable is calculated based on \( x \) number of days where \( x \) is a model parameter called rolling correlation days.
5. The historical data are filtered by plus or minus \( x \) days where \( x \) is a model parameter called rolling window days.
6. The top \( x \) number of independent variables most correlated to the dependent variable are identified where \( x \) is a model parameter called correlated variable number.
7. The independent variables identified in step 6 are scaled by the correlation score.
8. The filtered historical data is further segmented into weekdays, weekends, and holidays.
9. A block of hours in the filtered data is assembled based on the hour of the day being forecasted and the prior and posterior hours, the number of which are input parameters.
10. The Euclidean distance between the forecast block and every block of historical data is calculated and sorted.
11. The top \( x \) number of dates in the sorted list calculated in step 10 are used to identify and retrieve the scaled historical values of the dependent variable where \( x \) is a model parameter called sample size.
12. The mean of the historical values in the sample space is the forecast for that hour.
13. The difference between prior Horse estimates and actual observations are used to make a time series of model errors. These errors are passed into a univariate autoregression model to forecast the errors going forward. The results of this forecast are known as the Cart.
14. The Horse and Cart results are combined to produce the Horse and Cart forecast.
15. The forecasts are passed through a Savitzky-Golay filter to produce smoothed forecasts.

Exhibit 3 ERCOT Coastal Load Horse, Horse and Cart Forecast vs. Actual Load for July 5, 2019

Confidence Interval and Probability Score

In addition to the Horse and Cart forecast at every hour of the forecast duration, a confidence interval is calculated as well as a probability score. The confidence interval is simply the highest and lowest values in the sample space. A probability score is calculated using Laplace’s Rule of Succession, which is defined as

\[ P(A_i) = \frac{N_i + 1}{N + n} \]

where \( P(A_i) \) is the probability of the actual load falling into a given bin based on the number of past observations within the bin \( N_i \) divided by the total number of past observations \( N \) plus the total number of bins \( n \). If the sample size is greater than eight, the number of bins is eight plus two to account for the fact that, while no observations occurred outside the confidence interval, there is still a low probability for such an occurrence. If the sample size is less than eight, the sample space \( n \), or number of bins, is equal to the sample size plus two. In other words, the smaller the sample size, the more likely the actual observation will occur outside the previous observations in the sample. For example, if there are only three observations in the sample size, there is a \( (0+1)/(3+5) = 1/8 \) chance that the actual observation will be to the right or the left of the three observations in the forecast distribution. A probability score is associated with each bin. The midpoint of each of the middle bins is reported whereas the value of the inner edge of the outside bins are reported. Consider three observations in a sample with the values of 1, 4, and 4, respectively. The Horse forecast is 3. The confidence interval is 1 through 4. The probability distribution is 1/8, 2/8, 1/8, 3/8, and 1/8 for each of the bins in the sample space. The values associated with each bin are 1, 1.5, 2.5, 3.5, and 4, respectively. In other words, there is a 1/8 chance of the next observation being less than 1. There is a 2/8 chance, or 25% probability, of the next observation falling between 1 and 2, which is reported as the mid-point between the bin edges, or 1.5. The probability of the next observation falling between 2 and 3 is also 1/8 because, like the edge bins, there were no prior observations in this bin. Since there were two prior observations in the bin between 3 and 4, reported as 3.5, the probability of the next observation falling in this bin is the highest at 3/8. Finally, there is 1/8 probability that the next observation will be greater than 4. These results are summarized in Exhibit 4.

| Back-testing Report |

In the same spirit of model transparency with which the Horse and Cart framework was conceived, a back-testing report is available for each model, including a summary of past performance and input parameters. If there are changes to the model, including changes to input parameters, a back-testing report justifying such changes will be made available.
About Morningstar® Commodities Research™
Morningstar Commodities Research provides independent, fundamental research differentiated by a consistent focus on the competitive dynamics in worldwide commodities markets. This joint effort between Morningstar's Research and Commodities & Energy groups leverages the expertise of Morningstar's 23 energy, utilities, basic materials, and commodities analysts as well as Morningstar's extensive data platform. Morningstar Commodities Research currently focuses on North American power and natural gas markets with plans to expand coverage of other markets worldwide.

Morningstar, Inc. is a leading provider of independent investment research in North America, Europe, Australia, and Asia. The company offers an extensive line of products and services for individuals, financial advisors, and institutions. Morningstar’s Commodities & Energy group provides superior quality market data and analytical products for energy data management systems, financial and agricultural data management, historical analysis, trading, risk management, and forecasting.

For More Information
+1 800 546-9646 North America
+44 20 3194 1455 Europe
commoditydata-sales@morningstar.com

About Morningstar® Quantitative Research
Morningstar Quantitative Research is dedicated to developing innovative statistical models and data points, including the Morningstar Quantitative Rating, the Quantitative Equity Ratings and the Global Risk Model.

For More Information
+1 312 244-7541
lee.davidson@morningstar.com

22 West Washington Street
Chicago, IL 60602 USA
©2020 Morningstar. All Rights Reserved. Unless otherwise provided in a separate agreement, you may use this report only in the country in which its original distributor is based. The information, data, analyses, and opinions presented herein do not constitute investment advice; are provided solely for informational purposes and therefore are not an offer to buy or sell a security; and are not warranted to be correct, complete, or accurate. The opinions expressed are as of the date written and are subject to change without notice. Except as otherwise required by law, Morningstar shall not be responsible for any trading decisions, damages, or other losses resulting from, or related to, the information, data, analyses, or opinions or their use. The information contained herein is the proprietary property of Morningstar and may not be reproduced, in whole or in part, or used in any manner, without the prior written consent of Morningstar. To license the research, call +1 312 696-6869.