Finding Balance: Energy, Wealth, and Health

Anthony Forgetta

A Thesis
In the Department of
Mathematics and Statistics

Presented in Partial Fulfillment of the Requirements
for the Degree of
Doctor of Philosophy (Mathematics and Statistics) at
Concordia University
Montréal, Québec, Canada

May 2025

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Anthony Forgetta**

Entitled: *Finding Balance: Energy, Wealth, and Health*

and submitted in partial fulfillment of the requirements for the degree of

Doctor Of Philosophy (Mathematics and Statistics)

complies with the regulations of the University and meets the accepted standards with respect to originality and quality.

Signed by the final Examining Committee:

_____ Chair
Dr. E. Belilovsky

_____ External Examiner
Dr. T. Ware

_____ Examiner
Dr. Y. Chaubey

_____ Examiner
Dr. A. Sen

_____ Examiner
Dr. W. Sun

_____ Supervisor
Dr. F. Godin

Approved by _____ Dr. Cody Hyndman, Graduate Program Director

July 29, 2025 _____
Dr. Pascale Sicotte, Dean of Faculty

Abstract

Finding Balance: Energy, Wealth, and Health

Anthony Forgetta, Ph.D.
Concordia University, 2025

This dissertation is composed of three manuscripts (manuscript 1 has been published in Energy Economics, manuscript 2 is currently under review with the Journal of Energy Markets, and manuscript 3 is currently under review with the Obesity Research and Clinical Practice journal) that address problems in energy economics, finance, and epidemiology through statistical modeling.

The first manuscript proposes a covariate-dependent mixture model to describe the behavior of electricity DART spreads (defined as the difference between the day-ahead and real-time prices of electricity). The model incorporates multiple regimes and allows covariates to impact both the frequency and severity of DART spread spikes. Using data from the Long Island zone of the New York Independent System Operator, the model demonstrates a strong fit. Results reveal that including covariates in the severity component of the model is crucial, while mild additional performance is obtained with their inclusion in the frequency component. Neural network-based quantile regression benchmarks are unable to improve performance over our mixture model.

The second manuscript examines the diversification benefits of energy commodities during turbulent periods such as those marked by the COVID-19 pandemic and the Russia-Ukraine war, both of which deeply affected energy markets. Revisiting classical allocation strategies, we incorporate electricity futures—a rarely used asset—alongside crude oil and natural gas futures. Using mean-variance optimization, the diversification benefits are evaluated by combining these energy contracts with the S&P 500. Our empirical approach handles the non-stationarity of returns, volatilities, and correlations. Out-of-sample results show improved performance and diversification, especially during crisis periods.

The third manuscript extends existing dual-energy X-ray absorptiometry-based body composition classifications by introducing additional centile cut-offs to capture tail behavior. Using NHANES (National Health and Nutrition Examination Survey) data, we study the association between these phenotypes and health risks, including metabolic syndrome (MetS), depression, sleep disorders, and comorbidities. Nine phenotypes were identified using quantile regression (QR), and logistic regression was used to assess their relationship with health risks, compared to standard adiposity measures like body mass index (BMI), waist circumference (WC), and total fat percent. The QR model has a better (higher) LR+ (positive likelihood ratio) than the median-split model for MetS and comorbidity but consistently underperforms in LR- (negative likelihood ratio) compared to the median-split model. Both models perform worse than BMI and WC. Whether results differ over time or among certain subpopulations should be investigated.

Acknowledgments

I express my sincerest gratitude to my supervisor, Professor Frédéric Godin. He was my teacher and guide, and gave me the opportunity to realize my dreams. I would like to deeply thank my collaborators, Professor Maciej Augustyniak, Professor Geneviève Gauthier, and Professor Lisa Kakinami, who helped me mature personally, academically, and professionally. Your teachings and lessons will stay with me forever.

My thanks go to Concordia University for the knowledge and education that I gained during my time there.

I would like to thank Plant-E Corp and its team, in particular Pierre Plante, Vincent Marchal, Shahab Einabadi, and Elizabeth Fernandez de Ordonez, for providing access to their data, for their invaluable insight into the functioning of power markets, and for their support. In addition, I am grateful to MITACS for providing me with financial support and the opportunity to gain industry experience.

To my friends, thank you for all the happy memories, laughs, and smiles.

To my family, words cannot describe how I feel, they cannot do you justice. I will simply say this, thank you.

Contribution of Authors

This thesis is based on three research articles:

I: Forgetta, A., Godin, F., and Augustyniak, M. Distributional forecasting of electricity DART spreads with a covariate-dependent mixture model.

This joint work was accepted for publication in Energy Economics. Forgetta is responsible for a substantial portion of the analysis, as well as the primary portion of the writing with editing by Godin and Augustyniak.

II: Forgetta, A., Gauthier, G., and Godin, F. Do energy futures add value to stock portfolios?

This joint work has been submitted to the Journal of Energy Markets. Forgetta is responsible for a substantial portion of the analysis, as well as the primary portion of the writing with editing by Gauthier and Godin.

III: Forgetta, A. and Kakinami, L. On the use of quantile-regression-based DEXA phenotypes to assess health risks.

This joint work has been submitted to the Obesity Research and Clinical Practice journal. Forgetta is responsible for a substantial portion of the analysis, as well as the primary portion of the writing with editing by Kakinami.

Contents

List of Figures

List of Tables

1 Introduction

Across many disciplines, from energy and financial markets to public health, statistical modeling plays an important role in making decisions under uncertainty. Many real-world systems exhibit non-linear behaviors, non-Gaussian distributions, and regime shifts that complicate traditional methods. Whether the goal is to forecast price spikes, diversify portfolios with alternative investments, or classify people at risk of developing specific health issues, statistical modeling is essential to understanding these inherent complexities.

A common challenge in these applications is the ability to take into account structural changes and tail risks. As a use case, since electricity cannot be efficiently stored on a large scale, its supply and demand must always be in equilibrium. High-cost generators are often employed to meet unforeseen demand in real time, often leading to price spikes. Employing a covariate-dependent mixture model to capture both regular regimes and extreme spikes can therefore provide a framework for capturing these intricacies, as covariates can be directly incorporated into the model parameters.

In the case of portfolio management, the diversification potential of energy futures is similarly regime dependent; it is therefore influenced by crises such as the COVID-19 pandemic and the Russia-Ukraine war, and by events such as the financialization of commodity markets. In these cases, asset return models must, therefore, have the capacity to dynamically adapt to evolving market conditions and turbulent periods. A model that empirically forecasts expected asset returns and volatilities can, therefore, account for the structural breaks that occur during crises.

In epidemiology, obesity is an important area of continued investigation; the ability to model the entire distributions of muscle mass and fat mass can help reveal critical patterns hidden in the data. Quantile regression can help capture the full spectrum of data distributions, including tail behaviors, which have frequently been omitted by traditional models used in the literature on obesity.

From hedge funds to institutional investors to individuals, this collection of studies demonstrates that tail risks and extreme events must be understood to avoid significant financial costs. For power merchants, being able to adequately model underlying electricity price spikes can help prevent significant losses. For institutional investors seeking to diversify their portfolios, it is critical to understand the underlying asset price dynamics under different market regimes, such as the COVID-19 pandemic and the Russia-Ukraine war, for sound portfolio management. Even at an individual level, the ability to identify high-risk patients in the tails of the distribution, as with obesity, can potentially result in significant savings for the economy as a whole.

The remainder of this thesis is organized as follows. In the next section, chapter 2, we develop a covariate-dependent mixture model to forecast the distribution of hourly DART spreads. In our mixture model, the conditional distribution of the DART spread in the regular market regime is modeled using a Gaussian distribution, whereas spike regimes are modeled using Generalized Pareto distributions. Covariates such as forecasts of load, weather, and natural gas price, as well as cyclical indicators, are incorporated into the model parameters. Specifically, the parameters of all regime distributions, as well as the mixing probabilities, are expressed as functions of these covariates and are hence time-varying.

In chapter 3, we reexamine whether incorporating energy futures can add value to an equity portfolio by analyzing an updated data set that includes two new crises, namely the COVID-19 pandemic and the Russia-Ukraine war, which have significantly impacted the energy sector. Our

model empirically forecasts asset expected returns and volatilities and can adequately handle the non-stationary nature of the time series data. Asset allocation is performed using a mean-variance optimization approach, with the Sharpe ratio used as the objective function. We benchmark our performance against a static investment in the S&P 500.

In chapter 4, we develop a DEXA-derived (dual energy X-ray absorptiometry) phenotype classification system that captures kurtosis and considers additional centile cut-offs beyond the median. We then compare the performance of our models to standard adiposity measures, such as BMI (body mass index) and waist circumference. This study focuses on metabolic syndrome as the primary health outcome, but additionally assesses other health risks, such as depression, sleep disorders, general health, instrumental activities of daily living, and comorbidities.

At the end of this thesis, we provide an appendix that supplements the main chapters. It contains additional details that are mentioned within the text. The reader is directed to the appendix via references included throughout the main body of the text.

2 Distributional forecasting of electricity DART spreads with a covariate-dependent mixture model

2.1 Introduction

Electricity is a unique commodity as it cannot be efficiently stored at a large scale; that is, its supply and demand must always be in equilibrium. Transmission maintenance, unexpected interruptions, and unforeseen demand can give rise to transmission congestion, thus creating bottlenecks within the power grid. To meet the unforeseen demand in real-time, high marginal cost generators are often employed due to the slow ramp-up time of large, inexpensive power plants. Taken together, these characteristics give rise to price spikes (i.e., extreme values). An Independent System Operator (ISO) is thus needed to oversee the proper functioning of a power grid. For example, the New York ISO (NYISO) is the ISO responsible for the power grid covering the state of New York. The purchase and sale of electricity generally occurs in the day-ahead (DA) and real-time (RT) energy markets, at the corresponding DA and RT energy prices.

The DART spread, defined as the difference between the DA and RT prices, is a quantity of interest to several market participants. Wholesalers, for instance, perceive it as the opportunity cost of transacting in the DA market relative to the RT market. Since RT prices are more volatile than DA prices, wholesale power purchasers are willing to pay a premium to buy electricity in the DA market to avoid excessive price volatility. Indeed, Longstaff and Wang (2004) found that forward premia are embedded in DA electricity prices, reflecting the economic risks faced by market participants in the RT market, such as the possibility of extreme price spikes. Another group exposed to DART spreads includes virtual bidders, who must offset their DA transactions using the RT market. Since their profit is determined by the differential between DA and RT prices, modeling the DART spread is paramount to their success.

While the dynamics of DA or RT prices of electricity have been widely researched (see Weron, 2014, and the references therein), few studies have focused on modeling the DART spread. For instance, Das et al. (2022) use univariate time series models to generate point forecasts of the DART spread for the next hour, at a given node of the PJM (the Pennsylvania-New Jersey-Maryland Interconnection) market grid. Galarneau-Vincent et al. (2023) model the probability of a spike in the DART spread for every hour of the day for the Long Island zone in the NYISO; several covariates are used, including load, weather, and cyclical data (such as daily and yearly cycles in the DART spread).

To contribute to this literature, our study develops a covariate-dependent mixture model to forecast the distribution of hourly DART spreads. In our mixture model, the conditional distribution of the DART spread in the regular market regime is modeled using a Gaussian distribution, whereas spike regimes are modeled using Generalized Pareto distributions.[1] Covariates such as forecasts of load, weather, and natural gas price, as well as cyclical indicators, are incorporated into the model parameters. Specifically, the parameters of all regime distributions, as well as the mixing probabilities, are expressed as functions of these covariates and are hence time-varying. Our approach therefore extends the work by Davison et al. (2002), who model hourly electricity spot prices using a mixture of Gaussian distributions. As described in the survey paper by Weron (2014),

[1]The idea of using Generalized Pareto distributions to depict large electricity price movements has previously been explored in Klüppelberg et al. (2010).

most studies in the literature consider only a small set of covariates. Our study, however, employs a comprehensive set of covariates to extract as much information as possible and to determine their relative importance in explaining the DART spread. Moreover, we adopt the perspective of a market participant placing bids on both the DA and RT markets; as such, load forecasts, weather forecasts, and natural gas futures prices are used as they become available—our results therefore accurately reflect how this model would perform if used in real-time.

There exists a broad literature on the use of factor-based stochastic processes to generate electricity prices, see for instance Benth et al. (2007). This general approach, surveyed in Deschatre et al. (2021), can be used to generate distributional forecasts. Such stochastic factors framework has been developed with the objective of pricing derivatives, for instance. In the context of DART spread forecasting, our modeling framework has an important advantage over the stochastic factors approach: it circumvents the need to model the dynamics of a large number of covariates exhibiting heterogeneous behavior. Indeed, instead of constructing a joint distributional forecast for the set of all covariates and mapping such distribution into a future DART spread distribution, our approach directly constructs the distributional relationship between future DART spreads and current covariate values.

Other studies have computed distributional forecasts of either DA or RT prices using neural networks. Brusaferri et al. (2020) propose a mixture density network to generate distributional forecasts for electricity prices in the Italian DA energy market. Afrasiabi et al. (2022) implement a mixture density network to forecast the statistical properties of electricity prices in CAISO. Marcjasz et al. (2023) use distributional neural networks to generate probability forecasts for hourly DA electricity prices in the German market. These studies focus primarily on model performance, assessed through their ability to generate forecasts—as measured, for example, by the mean absolute error (MAE), root mean square error (RMSE), continuous ranked probability score, or Kupiec's test. Our study, however, analyzes the statistical fit of the proposed model, and assesses the importance of including a wide variety of covariates to explain the DART spread distribution. To determine whether the inclusion of covariates is needed in both the frequency and severity components of the mixture model, we run a nested model comparison. We also apply a Shapley decomposition to the log-likelihood to help identify the relative importance of each covariate in the model. To assess the predictive ability of our model upon unseen data, we conduct an out-of-sample experiment, wherein we examine whether regularization techniques can improve the out-of-sample performance of the model. Finally, an advantage to using our approach—in relation to neural networks—is increased model interpretability. This provides a significant advantage to market participants, for many of whom inference is just as relevant as the ability to generate accurate forecasts. Nevertheless, this study includes a comparison of the performance of our mixture model to that of neural-network-based quantile regression benchmarks to assess whether the better interpretability of our model leads to a sacrifice in accuracy.

Our findings suggest that the proposed covariate-dependent mixture model provides an adequate fit to the data. In-sample nested model comparisons highlight the importance of including the entire set of covariates, in both the frequency and severity components of the mixture model, although their inclusion in the severity component is more critical. In addition, the SAGE variable importance assessment algorithm indicates that on a relative basis, the natural gas futures price has the most significant effect upon model performance. This makes sense since fuel has a large influence on the marginal cost of producing power. The other groups of covariates have smaller, though non-negligible effects, each of which is similar in magnitude. Additionally, the regularization

methods are shown to improve the out-of-sample performance of our model. Finally, a comparison of our model to neural-network-based benchmarks inspired from the machine learning literature results in superior out-of-sample performance of our model on our test set in terms of DART spread quantile predictions.

The remainder of the chapter is organized as follows. Section 2.2 describes the dataset used for model construction and evaluation, which consists of the DART target variable and explanatory variables. Section 2.3 specifies the covariate-dependent mixture model. Section 2.4 presents the model estimation, its performance, and several robustness checks. Section 2.5 compares the performance of our model with that of a machine learning method performing quantile regression. Section 2.6 concludes.

2.2 Description of the data

This study aims to develop an effective methodology for generating distributional forecasts of hourly electricity DART spreads. Our approach is implemented using DART spread data for the NYISO Long Island zone. Long Island is a peninsula and is thus susceptible to extreme price volatility, making it an ideal candidate for testing risk management models. The DART spread at hour t is computed as

$$\mathrm{DART}_t = \mathrm{DA}_t - \mathrm{RT}_t,$$

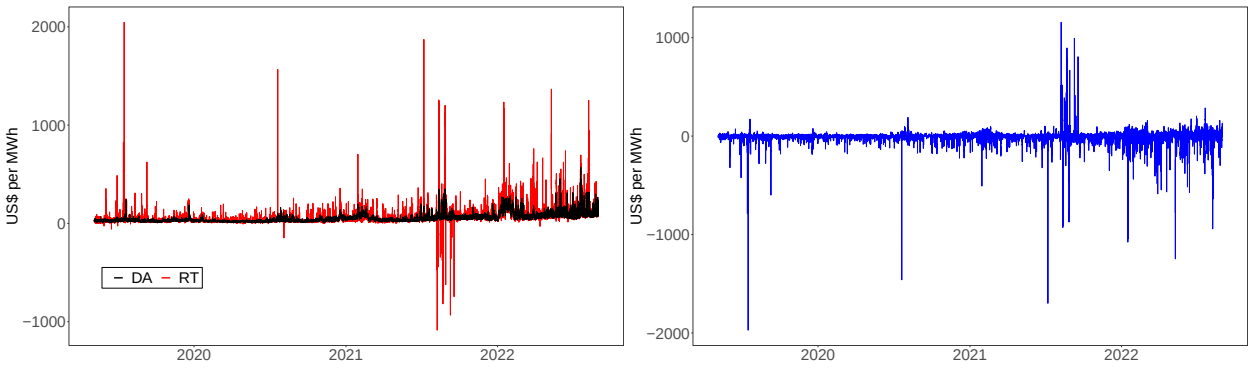
where DA_t and RT_t are respectively the DA and RT prices for hour t on a per MWh basis in US$. The NYISO's procedure for reporting prices is described in Section 2.2.1.

The dataset consists of 29172 hourly price observations spanning from April 26, 2019 12:00 to August 31, 2022 23:00.[2] Hours are expressed in the eastern time zone (i.e., EST) and we define the time index t such that the first hour in the dataset corresponds to $t = 0$.

Time series of the hourly DA and RT prices are displayed in Panel (a) of Figure 1, whereas Panel (b) provides the hourly DART time series. As expected, RT prices are more volatile than DA prices, with the presence of spikes being readily apparent. A significant portion of the dataset covers COVID-19 pandemic years (2020 and 2021) and the Russia-Ukraine war (starting February 2022). These events hold significant implications for power prices, arising from shifts in consumption patterns amidst the pandemic, further exacerbated by energy supply shortages during the conflict.

To assess the presence of cyclical patterns in the variability of the DART spread, we group the observed DART spreads into buckets based on either the hour of the day, month of the year, or day of the week. Figure 2 displays the mean, median, and 5% or 95% quantiles from the sample distribution, for each bucket. The observed time-of-day pattern in Panel (a) is caused by an increase in consumption during the on-peak hours occurring during the late afternoon. During these hours, the mean falls below the median, and the mean is consistently negative. This implies that the DA premium is not sufficiently compensating for the spikes that occur in RT in our sample. Furthermore, the quantiles increase in magnitude during these hours, thus indicating increased variability in the DART spread, which can be accounted for by the stress placed upon the grid due to the increased demand. The seasonality related to winter and summer seasons in Panel (b) is a consequence of increased electricity demand during these time periods stemming from more intensive heating

[2] Hours associated with daylight savings are discarded from the dataset.



(a) Hourly DA and RT prices.

(b) Hourly DART spread.

Figure 1: DA, RT, and DART hourly time series for the NYISO Long Island zone between April 26, 2019 and August 31, 2022.

and cooling requirements, respectively. The quantiles increase in magnitude during winter and summer months, once again due to the higher stress placed upon the grid as a result of the increased demand for power. Panel (c) does not indicate the presence of a strong seasonal effect related to the day of the week, both in level and in relation to the variability of the DART. These empirical observations lead us to include hourly and monthly seasonal covariates in our model, as they display clear seasonal patterns in the variability of the DART.

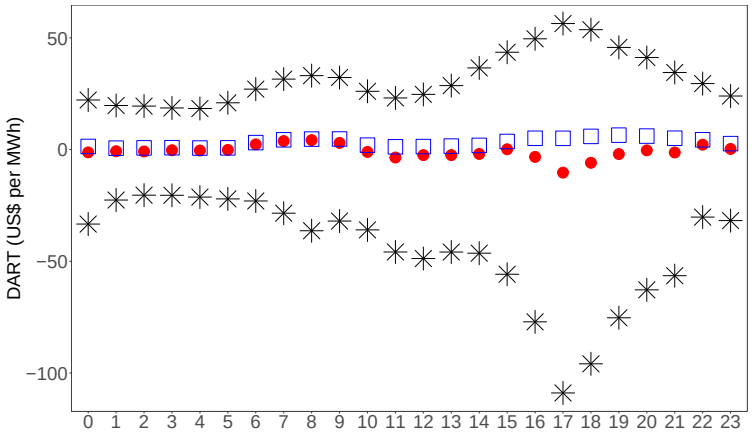
Descriptive statistics of the hourly DART spreads are provided in the first row of Table 1. Notice that the median and mean DART spread have opposite signs. This behaviour is systematically observed for each year in the sample, as indicated in Table 2. The positive median DART spread indicates that market participants are usually willing to pay a premium to lock-in a price for electricity on the next day. Indeed, the proportion of positive DART spreads in the data is 61%. However, as the mean DART spread is negative, the DART premium was insufficient on average to compensate for the large negative DART spikes observed in the sample.

2.2.1 Explanatory variables

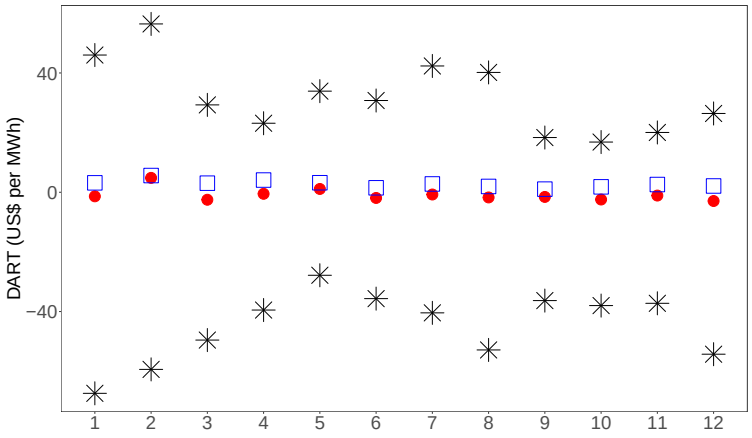
This section describes the variables that are used as covariates in our mixture model to generate distributional forecasts of hourly DART spreads. We consider herein the perspective of a market participant placing bids on both the DA and RT markets. The timeline to collect relevant information, place a set of bids, and have its outcomes revealed ranges over three days; these are referred to as the *information collection day*, the *trading day*, and the *target day*:

- *Information collection day* (Day 1): At 07:30 the NYISO publishes the load forecast for every hour of the *target day*. In addition, the two-day-ahead natural gas futures price is reported by the ICE.

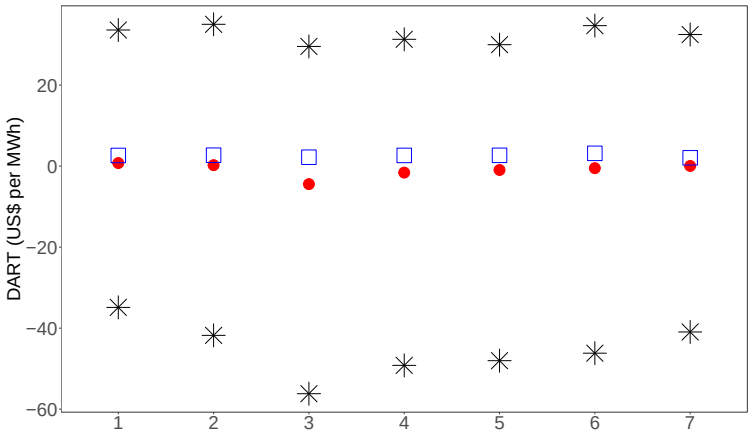
- *Trading day* (Day 2): At 04:00 the weather forecasts are published by the Darksky data provider for every hour of the *target day*. These forecasts are based upon measurements made by the Huntington weather station in New York. Furthermore, the market participants must



(a) Daily cycle, where 0 = 00:00,. . . , 23 = 23:00.



(b) Yearly cycle, where 1 = January,. . . , 12 = December.



(c) Weekly cycle, where 1 = Sunday,. . . , 7 = Saturday.

Figure 2: Daily, yearly, and weekly seasonal DART patterns.

Notes: Sample mean (red filled circles), median (blue squares), and 5% and 95% quantiles (black stars) of the DART spread, with Panel (a): hourly buckets, Panel (b): monthly buckets, Panel (c): daily buckets.

Table 1: Descriptive statistics for the DART spread and explanatory variables.

Name	Abbr.	Unit	Min	Q_1	Median	Mean	Q_3	Max	Std
DART spread	DART	US$/MWh	-1971.57	-4.57	2.61	-0.91	10.13	1155.56	46.03
Load forecast	LOAD	MWh	1281	1863	2154	2336	2586	5271	686.12
Heating Degree Day forecast	HDD	°F	0.00	0.00	4.00	6.67	12.55	30.72	7.34
Cooling Degree Day forecast	CDD	°F	0.00	0.00	0.00	2.10	3.78	17.78	3.36
Wind Speed forecast	WS	m/s	0.90	6.40	8.90	9.62	12.10	33.40	4.24
UV Index forecast	UV		0.00	0.00	0.00	1.18	2.00	10.00	1.96
Natural Gas Futures Price	NG	US$	2.12	3.98	5.04	7.00	9.21	26.32	4.35
Daily Fourier Basis (Sin)	SinD								
Daily Fourier Basis (Cos)	CosD								
Yearly Fourier Basis (Sin)	SinY								
Yearly Fourier Basis (Cos)	CosY								

Notes: The dataset ranges from April 26, 2019 12:00 to August 31, 2022 23:00.

Table 2: Mean and median of hourly DART spreads, by year.

Year	Mean	Median
2019	-2.29	1.48
2020	-0.06	2.57
2021	-0.86	2.27
2022	-0.86	6.53

place their bids for the DA NYISO auction by 05:00. At 11:00, the NYISO then publishes the DA prices for every hour of the *target day* based upon the latter auction results.

- *Target day* (Day 3): The RT prices are progressively revealed on an hourly basis throughout the day (from 00:00 to 23:00), allowing for the computation of the hourly DART spreads on the *target day*.

Therefore, every time a prediction is made for an hour of the *target day* d, explanatory variables related to the hourly weather forecasts are obtained on the *trading day* $d-1$, whereas the hourly load forecast and the natural gas price data are taken from the *information collection day* $d-2$. The same natural gas futures price is used as an explanatory variable for every hour t of the *target day*. The explanatory variables considered, along with their units of measurement and corresponding descriptive statistics, are presented in Table 1. These variables account for information pertaining to load forecasts, weather forecasts, natural gas futures prices, and cyclical indicators. In the literature, other studies have also used these covariates to forecast electricity prices, as described in Weron (2014). In most studies, only a subset of these covariates are considered. Our study, however, uses all these classes of covariates to extract as much information as possible. Such variables are now described in detail.

Load forecasts are included because high expected consumption can put upward pressure on electricity prices. Several variables are extracted from one-day-ahead weather-related forecasts. Such forecasts include temperature, wind speed, and UV index. Among these, temperature is not used directly, but is instead transformed into alternative covariates, namely, Heating Degree Day (HDD) and Cooling Degree Day (CDD), as follows:

$$\mathrm{HDD}_t = \max(\mathrm{BP} - \mathrm{T}_t, 0), \quad \mathrm{CDD}_t = \max(\mathrm{T}_t - \mathrm{BP}, 0), \tag{1}$$

where T_t is the temperature forecast for hour t and BP = 65 °F (18.3 °C) is the threshold recommended by the NYISO (Fox et al., 2019). Such transformations of the temperature forecasts are commonly used, including by the NYISO to generate load forecasts, see NYISO (2021) and Fox et al. (2019), and can more accurately capture the non-linearity present between temperature and electricity consumption. More precisely, a relatively larger amount of electricity is consumed when temperatures are either very high or very low due to a larger demand for cooling and heating, respectively. Alternatively, quadratic terms of the CDD and HDD might also be considered, but for the sake of parsimony, we only include linear terms.

The wind speed and UV index forecasts are used directly as explanatory variables. The wind speed has an effect on the amount of available wind generation, and thus upon the available supply of electricity. The UV index forecast is a proxy for the amount of available sunshine, which impacts the solar generation of electricity by suppliers. Moreover, available sunshine also affects the demand for electricity in the Long Island zone due to the presence of solar panels on residential and commercial buildings (referred to as *behind-the-meter generation*). The UV index takes on integer values between 0 and 10.

The main use for natural gas is in electricity generation, where it often serves as the marginal unit of production in the supply stack. Peaker plants, that is power generators used to quickly balance supply and demand in real time, are primarily fueled using natural gas. These features thus make it a significant driver of electricity prices. The natural gas futures price is therefore used as an explanatory variable.[3]

To account for the daily and yearly seasonalities in the distribution of the DART spread, as evidenced in Figure 2, the following variables are included in the model.[4] The daily cycle at hour t is represented by:

$$\mathrm{CosD}_t = \cos\left(\frac{2\pi t}{24}\right), \ \mathrm{SinD}_t = \sin\left(\frac{2\pi t}{24}\right), \tag{2}$$

and the yearly cycle at hour t by:

$$\mathrm{CosY}_t = \cos\left(\frac{2\pi t}{365 \times 24}\right), \ \mathrm{SinY}_t = \sin\left(\frac{2\pi t}{365 \times 24}\right). \tag{3}$$

Table 3 displays Pearson correlations between the explanatory variables used in the mixture model. The CDD-LOAD pair exhibits a strong positive correlation, since as the demand for cooling

[3]The natural gas delivery hub is Transco Zone 6 (NY), with an underlying of 2500 million British thermal units (i.e., 2500 MMBtus).

[4]In unreported experiments, we also tested the use of dummy variables instead of sine and cosine functions to represent seasonality, e.g. a dummy variable for each month or each hour of the day. Models using dummies were generally not able to significantly improve performance, while substantially increasing computational cost.

Table 3: Sample Pearson correlations between explanatory variables.

	LOAD	WS	UV	HDD	CDD	NG
LOAD	1					
WS	-0.12	1				
UV	0.38	-0.06	1			
HDD	-0.38	0.23	-0.32	1		
CDD	0.88	-0.19	0.49	-0.57	1	
NG	0.07	0.00	-0.01	0.11	0.03	1

Notes: The dataset ranges from April 26, 2019 12:00 to August 31, 2022 23:00.

increases, the corresponding load also increases. A few other pairs of variables possess statistically significant linear relationships, such as LOAD-UV, LOAD-HDD, UV-CDD, and HDD-CDD; however, their correlations are not strong enough for the variables to be redundant, and as such none of these are excluded from the model. In addition, Figure 3 provides a scatterplot illustrating the relationships between the explanatory variables in the sample. It is clear that highly non-linear and non-trivial relationships exist between the remaining pairs of covariates.

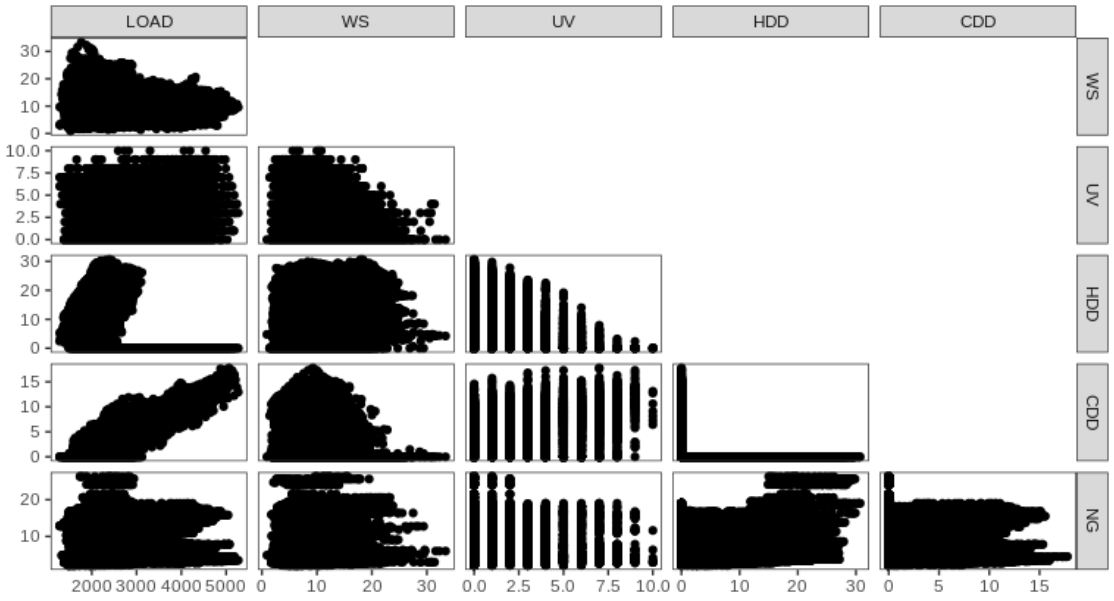


Figure 3: Scatterplot of the explanatory variables.

2.3 Model specification

This section proposes to model the DART spread distribution for every hour of the *target day* with a three-regime mixture process that incorporates the covariates discussed in Section 2.2.1. Our model allows for time-varying mixing probabilities, and it can reproduce the following salient features

of DART spreads: (1) the occurrence of large negative and positive spikes, (2) the relationships between DART spread volatility and temperature or energy-related covariates, and (3) the presence of seasonal patterns over daily and yearly cycles as discussed in Section 2.2.

Let y_t and $\mathbf{x}_t$ denote, respectively, the DART spread at hour t and the vector of covariates used to predict the DART spread distribution for that hour. The covariates considered are those from Table 1, that is:

$$\mathbf{x}_t = [1,\ \mathrm{LOAD}_t,\ \mathrm{HDD}_t,\ \mathrm{CDD}_t,\ \mathrm{WS}_t,\ \mathrm{UV}_t,\ \mathrm{CosD}_t,\ \mathrm{SinD}_t,\ \mathrm{CosY}_t,\ \mathrm{SinY}_t,\ \mathrm{NG}_t]. \tag{4}$$

Note that as $\mathbf{x}_t$ is used to explain the DART spread at hour t of the *target day*, it contains information that is provided either on the *information collection day* or the *trading day*.

The mixture model is composed of three distinct regimes for the DART spread: a regular regime, a positive spike regime, and a negative spike regime. The probability density function (PDF) of the mixture model is given by:

$$f(y_t|\mathbf{x}_t) = \sum_{i=1}^{3} \pi_i(\mathbf{x}_t) f_i(y_t|\mathbf{x}_t), \quad y_t \in \mathbb{R},$$

where π_i and f_i are, respectively, the covariate-dependent mixing probability and conditional PDF for regime i ($i = 1,2,3$). A multinomial logistic mapping is used to model the mixing probabilities as a function of the covariates, that is,

$$\pi_i(\mathbf{x}_t) = \frac{\exp^{(\mathbf{x}_t^\intercal \boldsymbol{\beta_i})}}{1 + \sum_{k=2}^{3} \exp^{(\mathbf{x}_t^\intercal \boldsymbol{\beta_k})}}, \quad i = 2,3, \tag{5}$$

and $\pi_1(\mathbf{x}_t) = 1 - \pi_2(\mathbf{x}_t) - \pi_3(\mathbf{x}_t)$.

The conditional distribution of the DART spread in the first regime at hour t (i.e., $f_1(y_t|\mathbf{x}_t)$) is assumed to be Gaussian with mean $\mu_1(\mathbf{x}_t)$ and variance $\sigma_1^2(\mathbf{x}_t)$. This normal regime is intended to produce DART realizations in *regular* days (i.e., in days when the electricity market is not subject to excessive stress). The conditional mean in this regime is specified as a linear combination of covariates: $\mu_1(\mathbf{x}_t) = \mathbf{x}_t^\intercal \boldsymbol{\eta_1}$, whereas the conditional scale parameter, $\sigma_1(\mathbf{x}_t)$, is set to $\sigma_1(\mathbf{x}_t) = h(\mathbf{x}_t^\intercal \boldsymbol{\zeta_1})$, where h is a function[5] that transforms the linear predictor $\mathbf{x}_t^\intercal \boldsymbol{\zeta_1}$. More specifically, h is set to $h(x) = 10 \arctan x + 5\pi$, with domain $\mathbb{R}$ and range $(0,\ 10\pi)$. The arctan transformation in the specification of h ensures a positive and bounded conditional variance. The upper bound, set at 10π, is sufficiently large so that it is reached only for a very limited number of observations within the sample.

Since DART spreads can exhibit very large spikes, we model the spike regimes using the Generalized Pareto Distribution (GPD), which is well-suited for capturing heavy-tailed behavior. The GPD is often employed to model tail distributions due to the result of Pickands III (1975), which establishes that the tails of a wide class of distributions converges asymptotically to the GPD. Moreover, since the tail distribution of a GPD remains GPD-distributed (with modified parameters), assuming a GPD distribution for the DART spread in spike regimes produces GPD-distributed tails in such regimes. States 2 and 3 are intended to create regimes that can allow for large positive and

[5]Empirically, we found that placing an upper bound on the variance terms improved statistical fit. The upper bound was optimized through manual hyperparameter tuning.

negative DART spikes, respectively. Their conditional distributions are assumed to be:

$$f_2(y_t|\mathbf{x}_t) = f_{\text{GPD}}(y_t;\xi_2,\sigma_2(\mathbf{x}_t)), \quad f_3(y_t|\mathbf{x}_t) = f_{\text{GPD}}(-y_t;\xi_3,\sigma_3(\mathbf{x}_t)).$$

where f_{GPD} denotes the PDF of the GPD:

$$f_{\text{GPD}}(y;\xi,\sigma) = \frac{\left(1+\xi\frac{y}{\sigma}\right)^{\left(\frac{-1}{\xi}-1\right)}}{\sigma}\mathbb{1}_{\{y\geq 0\}}, \tag{6}$$

with scale and shape parameters σ and ξ, respectively. The third regime corresponds to a reflection of the GPD to allow for negative values. When the shape parameter satisfies $\xi > 0$, the GPD is considered a heavy-tailed distribution, thus making it an ideal candidate to generate spikes. The ξ_2 and ξ_3 parameters are constrained to the interval $(0,\ \frac{1}{2})$ to ensure that the GPD has a finite mean and variance. Similarly to the Gaussian regime, the conditional scale parameters in the spike regimes are obtained through a transform, h, applied to a linear combination of covariates: $\sigma_i(\mathbf{x}_t) = h(\mathbf{x}_t^\intercal \boldsymbol{\zeta}_i)$, for $i = 2,3$ (the same transform h is used as in the Gaussian specification). Since the variance of the GPD described in (6) is given by $\frac{\sigma^2}{(1-\xi)^2(1-2\xi)}$, the conditional variance in the spike regimes can be orders of magnitude greater than $\sigma_i^2(\mathbf{x}_t)$. For example, if $\xi_i = 0.42$, which is the value that we estimate for the negative spike regime (see Section 2.4), the conditional variance is almost 20 times the value of $\sigma_i^2(\mathbf{x}_t)$. We remark that the realizations obtained from the spike regimes are not always spikes per se. Indeed, since the support of the distributions in regimes 2 and 3 are, respectively, $[0,\infty)$ and $(\infty,0]$, these regimes can also produce mild DART values. However, the likelihood of producing a value that is many standard deviations away from the mean is significantly larger in these regimes than in the Gaussian one. As a consequence, π_2 and π_3 should not be understood directly as spike probabilities, although they influence the frequency of large DART outcomes.

Finally, the overall mean and variance of the DART spread in our proposed mixture model are respectively given by:

$$\mathbb{E}[y_t|\mathbf{x}_t] = \pi_1\mu_1 + \pi_2\frac{\sigma_2}{1-\xi_2} - \pi_3\frac{\sigma_3}{1-\xi_3},$$

$$\mathbb{V}[y_t|\mathbf{x}_t] = \pi_1(\sigma_1^2+\mu_1^2) + \pi_2\Big(\frac{\sigma_2^2}{(1-\xi_2^2)(1-2\xi_2)} + \big(\tfrac{\sigma_2}{1-\xi_2}\big)^2\Big) + \pi_3\Big(\frac{\sigma_3^2}{(1-\xi_3^2)(1-2\xi_3)} + \big(\tfrac{\sigma_3}{1-\xi_3}\big)^2\Big) - \big(\mathbb{E}[y_t|\mathbf{x}_t]\big)^2.$$

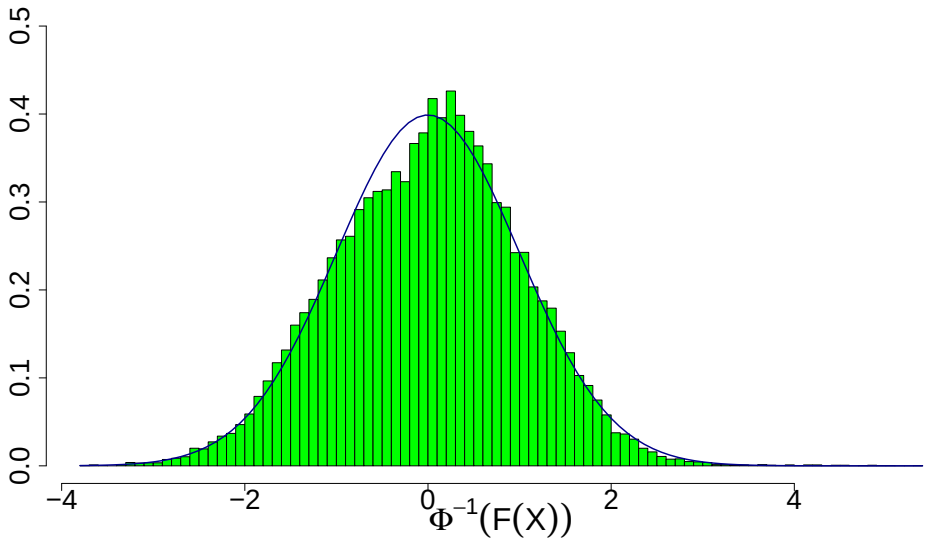
2.4 Model performance

This section discusses the model estimation results, the diagnostic tests performed to assess model adequacy, the relative importance of covariates and out-of-sample predictive tests.

2.4.1 Estimated parameters and goodness-of-fit

Model estimation is performed using maximum likelihood on the data set described in Section 2.2, which contains hourly data for the DART spread, as well as corresponding explanatory variables, between April 2019 and August 2022.[6]

[6]The `solnp` package in `R` is used to perform the minimization numerically.



Notes: This plot is used to visually assess for normality. Descriptive statistics for the normal Rosenblatt residuals are given by: sample mean = 0.004, sample standard deviation = 1.010, sample skewness = -0.019, and sample excess kurtosis = 0.060.

Figure 4: Rosenblatt transform.

Prior to fitting the model, explanatory variables are standardized to a common scale. For any hour-t, the standardized value of the i^{th} explanatory variable is denoted by $\tilde{x}_t^{(i)}$ and defined as:

$$\tilde{x}_t^{(i)} = \frac{x_t^{(i)} - \bar{x}^{(i)}}{s^{(i)}}, \tag{7}$$

where $\bar{x}^{(i)}$ and $s^{(i)}$ denote, respectively, the sample mean and sample standard deviation of $x_t^{(i)}$.

To verify our modeling assumptions from a statistical perspective, we perform an analysis of the Rosenblatt residuals, a metric that allows to gauge model adequacy at the distributional level, rather than simply in terms of point forecasts. Denote by $\theta = [\boldsymbol{\eta_1}, \boldsymbol{\zeta_1}, \boldsymbol{\zeta_2}, \boldsymbol{\zeta_3}, \boldsymbol{\beta_1}, \boldsymbol{\beta_2}, \xi_2, \xi_3]$ the set of model parameters and by $\hat{\theta}$ its maximum likelihood estimate. The normal Rosenblatt residual is given by $\Phi^{-1}(F(y_t \mid \mathbf{x}_t; \hat{\theta}))$, where $F(\cdot \mid \mathbf{x}_t; \theta)$ corresponds to the CDF of the mixture model and Φ^{-1} is the inverse CDF of a standard normal random variable. If the model is well-specified, then $\Phi^{-1}(F(y_t \mid \mathbf{x}_t; \hat{\theta}))$ should be approximately standard normal. Figure 4 shows a density histogram of the Rosenblatt residuals computed from our model. From a visual standpoint, the model appears to provide a good fit to our dataset. In addition, the Jarque-Bera normality test (see Jarque and Bera (1987)) applied to the normal Rosenblatt residuals does not reject the null hypothesis at the 1% significance level (i.e., the p-value is 4.69%), thus providing further evidence in favor of the model.[7]

[7]The Jarque-Bera test has very high power due to the large number of observations, making it particularly sensitive to even minor deviations from normality. Since the sample skewness and excess kurtosis of the normal Rosenblatt residuals are close to zero, we consider the outcome of the statistical test acceptable in our context.

To determine whether the inclusion of covariates is needed in both the frequency and severity components of the mixture model, we compare the log-likelihoods of three nested models, in addition to the full model. The results of this comparison are shown in Table 4. The *null-freq-null-sev* model does not depend upon any covariates. The *full-freq-null-sev* model only integrates covariates into the frequency components (i.e., into the mixing probabilities π_i, $i = 1,2,3$), whereas the *null-freq-full-sev* assumes that mixing probabilities are constant but allows the parameters of the severity components (i.e., regimes) μ_1 and σ_i, $i = 1,2,3$, to depend on covariates. The *full-freq-full-sev* model refers to the full model outlined in Section 2.3. In starting with the full model, the removal of covariates from the severity component leads to a relatively large increase in the BIC (from 242,412 to 246,111). Conversely, the removal of covariates from the frequency component does not result in as large of an increase (from 242,412 to 242,643). These findings, along with the likelihood ratio p-values comparing the full model and nested versions, suggest that the inclusion of covariates in both the frequency and severity components improves the fit of the mixture model.

Table 4: Full sample nested model comparison.

Model	Params	BIC	Log-likelihood	Likelihood ratio p-value
null-freq-null-sev	8	252601	-126259	< 0.01
full-freq-null-sev	28	246111	-122912	< 0.01
null-freq-full-sev	48	242643	-121075	< 0.01
full-freq-full-sev	68	242412	-120856	

Notes: For the likelihood ratio test, nested models are compared to the full model structure (i.e., full-freq-full-sev), with $H_0 =$ nested model is superior.

Table 5 provides estimated parameters for the model, along with the corresponding standard errors in parentheses. When all standardized covariates are set to 0, the probabilities of occurrence of the positive and negative spike regimes (states 2 and 3) are 0.21 and 0.32, respectively.[8] In addition, for both the frequency and severity components, a large majority of the loadings are statistically significant, in the sense that their associated asymptotic Gaussian confidence intervals do not include zero. For any given covariate, the $\hat{\beta}_2$ and $\hat{\beta}_3$ loadings consistently have the same sign, implying that spike regime probabilities either both increase or both decrease simultaneously when the corresponding covariate moves. For example, for positive loadings, increasing the covariate relative to its mean value causes the underlying DART distribution to become more volatile and increases the likelihood of observing a spike (and conversely for negative loadings). Furthermore, since $\hat{\xi}_3 > \hat{\xi}_2$, the left tail of the DART spread in our model is heavier than the right tail. This is consistent with our empirical observations in Panel (b) of Figure 1, where negative spikes are larger in absolute terms relative to positive spikes.

Moreover, since all parameter loadings corresponding to the LOAD covariate are positive, spikes are more likely and tend to be larger in magnitude when the LOAD is high. The fact that the LOAD has an important influence upon spikes is expected. Indeed, a large demand for electricity tends to generate stress on the grid, thus creating greater uncertainty. Similarly, every parameter loading

[8] Setting all standardized covariates to 0 is equivalent to setting all predictors to their sample mean. The regime probabilities are then computed using Equation (5).

associated with the HDD covariate is positive, indicating that high values of HDD contribute to an increased volatility in the DART distribution. However, for the CDD and NG covariates, the frequency and severity loadings have opposite signs for each one of the two spike regimes (i.e., regimes 2 and 3). The net effect on the DART distribution attributed to an increase in CDD or NG is therefore unclear when simply looking at parameter values.

To gain additional insight into the net impact of the covariates upon the DART distribution, we present a sensitivity analysis in Table 6. Starting from their average value of 0, we shock each covariate individually by one standard deviation in either direction, noting its effect upon the mean, standard deviation, and lower as well as upper quantiles (at the 1% and 99% levels, respectively) of the DART distribution. As expected based upon our previous discussion of Table 5, increasing either the LOAD or HDD leads to a corresponding increase in the variability of the DART. Amongst the weather covariates, HDD has the largest effect upon the DART variability; however, the impact of CDD is rather marginal. This finding suggests that winter months, which are associated with a greater demand for heating, lead to greater variability in the DART relative to summer months. Furthermore, decreasing the UV leads to an increase in the DART variability. This is intuitive, since a smaller value for UV reduces the supply of inexpensive energy, thus increasing uncertainty[9]. Although it is difficult to specify the net impact of NG upon the DART using Table 5, it is clear from Table 6 that higher levels of NG lead to a net increase in the variability of the DART. This too agrees with intuition, since as NG becomes more expensive, the uncertainty in the DART increases. This parallels our discussion in Section 2.2.1, where we explain that natural gas peakers often serve as the marginal unit of production in the supply stack.

Table 5: Model parameter estimates.

	Frequency		**Severity**					
Covariate	$\hat{\boldsymbol{\beta}}_2$	$\hat{\boldsymbol{\beta}}_3$	$\hat{\boldsymbol{\eta}}_1$	$\hat{\boldsymbol{\zeta}}_1$	$\hat{\boldsymbol{\zeta}}_2$	$\hat{\boldsymbol{\zeta}}_3$	$\hat{\xi}_2$	$\hat{\xi}_3$
Intercept	-0.80 (0.07)	-0.38 (0.05)	8.83 (0.20)	-1.05 (0.03)	-0.65 (0.08)	0.08 (0.05)	0.18 (0.01)	0.42 (0.01)
LOAD	0.39 (0.15)	0.25 (0.10)	1.14 (0.29)	0.27 (0.06)	1.30 (0.13)	0.58 (0.12)		
HDD	0.25 (0.11)	0.09 (0.08)	0.29 (0.28)	0.06 (0.07)	0.97 (0.09)	0.48 (0.08)		
CDD	0.72 (0.15)	0.41 (0.11)	1.07 (0.31)	0.20 (0.07)	-0.37 (0.15)	-0.14 (0.12)		
WS	0.00 (0.04)	0.06 (0.03)	0.40 (0.10)	-0.01 (0.02)	0.07 (0.03)	0.10 (0.03)		
UV	-0.17 (0.08)	-0.12 (0.06)	-1.73 (0.18)	0.05 (0.04)	-0.42 (0.06)	-0.31 (0.07)		
SinD	-1.11 (0.13)	-0.64 (0.09)	-2.72 (0.30)	-0.35 (0.07)	0.70 (0.11)	0.52 (0.10)		
CosD	-0.23 (0.05)	-0.30 (0.04)	-0.76 (0.10)	0.06 (0.02)	0.37 (0.05)	0.21 (0.04)		
SinY	0.01 (0.08)	0.07 (0.06)	2.48 (0.19)	0.15 (0.04)	0.15 (0.06)	0.68 (0.07)		
CosY	0.18 (0.06)	0.18 (0.05)	1.17 (0.12)	-0.08 (0.03)	0.18 (0.04)	-0.02 (0.04)		
NG	-0.20 (0.05)	-0.13 (0.04)	1.98 (0.15)	0.71 (0.02)	1.71 (0.09)	1.19 (0.06)		

Notes: Estimated parameters for the model described in Section 2.3, along with the corresponding standard errors in parentheses.

Next, we consider the time evolution of the mixture model parameters. That is, for every hour t, the model uses the covariate vector $\mathbf{x}_t$ to generate the following parameter estimates: $\hat{\mu}_1$, $\hat{\pi}_i$, $\hat{\sigma}_i$, $i = 1,2,3$. Figure 5 reports the evolution of all such parameters using time series plots, which display clear cyclical patterns. The red shaded regions correspond to summer months (April

[9]This is especially true for the Long Island zone, for which, as previously mentioned, solar energy plays an important role due to the large presence of residential and commercial solar panels (i.e. *behind-the-meter generation*). Whether this also holds for other zones that are less dependent on solar energy would need to be investigated.

Table 6: Effects of covariate shocks upon the DART distribution.

	Mean			**Std. Dev.**			**VaR**(1%)			**VaR**(99%)		
Covariate	-1	0	+1	-1	0	+1	-1	0	+1	-1	0	+1
LOAD	-0.37	-2.37	-1.98	20.75	31.79	45.42	-81.72	-128.30	-174.09	23.00	40.16	94.81
HDD	-0.93	-2.37	-1.88	23.16	31.79	41.79	-91.92	-128.30	-162.10	26.09	40.16	81.38
CDD	-1.57	-2.37	-2.76	31.71	31.79	30.37	-126.40	-128.30	-123.25	42.11	40.16	37.21
WS	-1.64	-2.37	-3.16	29.42	31.79	34.20	-117.46	-128.30	-139.37	38.50	40.16	41.93
UV	-2.76	-2.37	-1.73	38.36	31.79	25.53	-154.30	-128.30	-102.09	56.24	40.16	30.29
NG	-0.10	-2.37	-2.03	15.02	31.79	48.60	-58.24	-128.30	-187.30	19.55	40.16	91.48

Notes: Starting from an average value of 0, we move each standardized covariate individually one standard deviation in either direction, noting its effect upon the mean, standard deviation, and lower as well as upper quantiles (at the 1% and 99% levels, respectively) of the DART distribution.

through September), whereas the blue shaded regions correspond to winter months (October through March). In each panel, the black center line represents a weekly rolling average (i.e., 168 hourly observations), whereas the grey outline corresponds to the actual hourly observations. On top of the dynamic mixture parameters, the time evolution of several conditional mixture distribution metrics (median, mean, tail percentiles) are also reported.

Panel (a) in Figure 5, which depicts $\hat{\pi}_1$, indicates that the regular DART regime is more likely to occur in summer than in winter. This is consistent with Panels (b) and (c), respectively displaying $\hat{\pi}_2$ and $\hat{\pi}_3$, which suggest that spikes are more likely to occur in winter than in summer. Panels (k) and (l), which exhibit $\widehat{\text{VaR}}(1\%)$ and $\widehat{\text{VaR}}(99\%)$ respectively, confirm this observation. Moreover, since the conditional mean of the DART spread, $\hat{\mathbb{E}}(y_t \mid \mathbf{x}_t)$, depicted in Panel (i), peaks during the winter season, the DART spread is generally expected to be larger during this period. This agrees with our previous findings, in which the demand for heating (HDD) was seen to have a much larger impact on the variability of the DART spread than the demand for cooling (CDD). Once again looking at $\hat{\mathbb{E}}(y_t|\mathbf{x}_t)$ in Panel (i), we do not observe extreme outcomes being generated. In fact, the rolling average of the latter quantity oscillates within a band defined by the first and third quartiles of the empirical distribution of the DART spread (see Table 1). Of special interest is the median DART spread displayed in Panel (h). The positive median DART spread indicates that market participants are willing to pay a premium to lock-in a price for electricity on the next day. The mean DART spread in Panel (i), however, is often negative. This indicates that our model suggests that the DART premium is insufficient to compensate for the large negative DART spikes. These results are consistent with the stylized facts observed in Table 2. Finally, the rise in the magnitude of $\hat{\sigma}_1$, $\hat{\sigma}_2$, $\hat{\sigma}_3$, $\hat{\mu}_1$, $\widehat{\text{VaR}}(50\%)$, $\hat{\mathbb{V}}(y_t|\mathbf{x}_t)$, $\widehat{\text{VaR}}(1\%)$, and $\widehat{\text{VaR}}(99\%)$ that occurs in 2022— see Panels (d), (e), (f), (g), (h), (j), (k), and (l), respectively—is driven by the corresponding steep increase in the average natural gas futures price, displayed in Figure 6.

Figure 7 assesses the relationship between the spike regime probabilities. This analysis aims to identify the presence of asymmetrical risk (i.e., skewness) in the regime probabilities across winter and summer months. For example, Panels (a) and (b) in Figure 7 indicate that $\hat{\pi}_2$ and $\hat{\pi}_3$ are correlated in summer, but not in winter; Panels (b) and (c) in Figure 5, which respectively present $\hat{\pi}_2$ and $\hat{\pi}_3$, further corroborate this finding. This suggests that one spike regime dominates the other in winter, but not in summer. This leads to an asymmetrical regime probability risk in winter, but a more symmetrical risk in summer.

2.4.2 Variable importance assessment

To help identify the relative importance of each covariate in the model, the SAGE algorithm is considered (Covert et al., 2020). It applies a Shapley decomposition to the performance metric—in our case, the log-likelihood—and thus decomposes total predictive performance into a sum of contributions from the individual covariates. The SAGE algorithm has the favorable property of being applicable for any model producing predictive distributions. The contribution of a given covariate c to the hour-t log-density is given by:

$$\phi_{c,t} = \sum_{S \subseteq C \setminus \{c\}} \frac{|S|!(|C| - |S| - 1)!}{|C|!} \left[l_{S \cup \{c\}}(\mathbf{x}_{t,S \cup \{c\}}) - l_S(\mathbf{x}_{t,S}) \right],$$

where C is the set of all covariates, $|\cdot|$ denotes the cardinality of a set, $\mathbf{x}_{t,S}$ corresponds to the hour-t covariate vector for the subset of covariates S, and $l_S(\mathbf{x}_{t,S})$ is the hour-t log-density of the mixture model trained exclusively using the subset of covariates S. This can be shown to lead to

$$l_C(\mathbf{x}_{t,C}) = \phi_{\emptyset,t} + \sum_{c \in C} \phi_{c,t},$$

implying that the total outperformance over the trivial model (i.e., without covariates) is fully allocated to the various covariates. Moreover, the relative importance of covariate c can be measured by

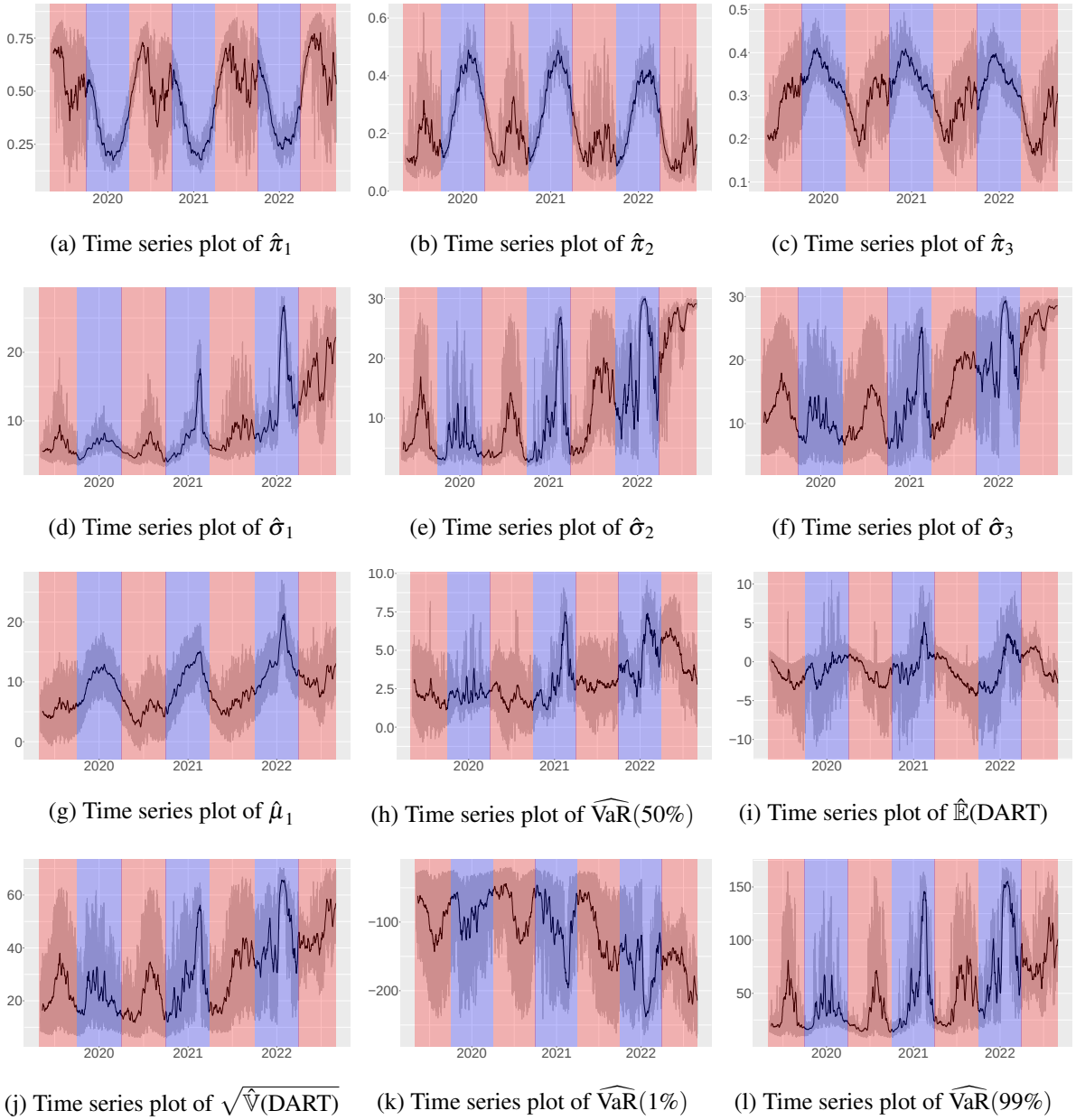
$$\psi_c = \sum_{t=1}^{\tau} \phi_{c,t},$$

where τ is the number of hours in the sample. Larger values of ψ_c indicate greater relative importance of covariate c. The SAGE algorithm can thus be used to assess the incremental effect of each covariate (or group of covariates) upon the log-likelihood. We thus use the Shapley decomposition to evaluate the relative importance of the following groups of covariates: **load** = [LOAD], **weather** = [HDD, CDD, WS, UV], **cycle** = [SinD, CosD, SinY, CosY], and **fuel** = [NG]. Table 7 displays the contributions to the log-likelihood from each group of covariates. **Fuel** has the largest contribution on a relative basis, and thus helps to explain the dynamics of the DART distribution to a greater extent than the other covariates. The other groups of covariates have smaller, though non-negligible effects.

Table 7: Covariate group contributions to the log-likelihood based on the SAGE algorithm.

Group of covariates c	Shapley importance score ψ_c	Shapley relative importance
Fuel	2367.27	0.44
Cycle	1274.29	0.24
Weather	956.35	0.18
Load	804.98	0.15

Notes: The Shapley relative importance is calculated by dividing the Shapley importance scores, ψ_c, by their total sum across all covariate groups.



(a) Time series plot of $\hat{\pi}_1$ (b) Time series plot of $\hat{\pi}_2$ (c) Time series plot of $\hat{\pi}_3$

(d) Time series plot of $\hat{\sigma}_1$ (e) Time series plot of $\hat{\sigma}_2$ (f) Time series plot of $\hat{\sigma}_3$

(g) Time series plot of $\hat{\mu}_1$ (h) Time series plot of $\widehat{\text{VaR}}(50\%)$ (i) Time series plot of $\hat{\mathbb{E}}(\text{DART})$

(j) Time series plot of $\sqrt{\hat{\mathbb{V}}(\text{DART})}$ (k) Time series plot of $\widehat{\text{VaR}}(1\%)$ (l) Time series plot of $\widehat{\text{VaR}}(99\%)$

Figure 5: Time series of the mixture model parameter estimates and corresponding mixture distribution risk metrics.

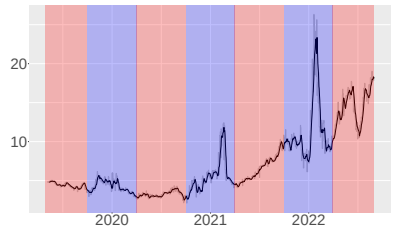
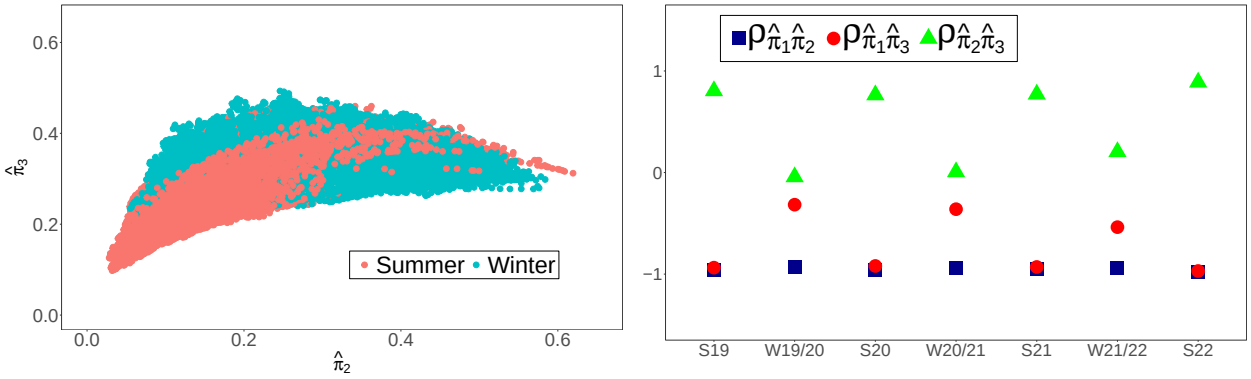


Figure 6: Time series plot of the natural gas price NG.



(a) Scatterplot between $\hat{\pi}_2$ and $\hat{\pi}_3$.

(b) Pairwise Pearson correlations between $\hat{\pi}_1$, $\hat{\pi}_2$, and $\hat{\pi}_3$ for various time periods. S = Summer months (i.e., April through September), W = Winter months (i.e., October through March), 19 = 2019, 20 = 2020, 21 = 2021, 22 = 2022.

Figure 7: Relationships between regime frequencies.

2.4.3 Out-of-sample performance

To assess the predictive ability of our model upon unseen data (i.e., data not used to train the model), we conduct an out-of-sample experiment wherein we examine whether regularization techniques—log-likelihood penalization according to the magnitude of parameters—can improve the out-of-sample performance of the model. Having a large number of parameters can lead to overfitting, which can deteriorate out-of-sample performance. Regularization techniques can thus be used to mitigate the potential impact of overfitting. Optimal parameters are therefore obtained by minimizing the following objective function:

$$\mathscr{PL}(\lambda,\theta) = -\left(\sum_{t=1}^{\tau} \log f(y_t \mid \mathbf{x}_t;\theta)\right) + \frac{1}{2}\lambda \left\|\theta\right\|_1^1, \tag{8}$$

where $\left\|\theta\right\|_1^1$ denotes the L1-norm, or LASSO-type penalty.

The hyperparameter λ is tuned using a grid search and sequential validation approach. More precisely, a value for λ is initially selected amongst the following set:[10]

$$\Lambda = \{0, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000\}.$$

Then, to perform the sequential validation, at the start of every month between May 1, 2020 and August 1, 2022 inclusively, the model is re-trained using all available historical data; the newly calibrated parameters are then used to make forecasts for that month. For example, on July 1, 2021 the model is re-trained using all available data between April 26, 2019 12:00 (i.e., the start of the dataset) and June 29, 2021 23:00 inclusively; we label this training data as $\mathscr{D}$.[11] These parameters

[10] Separate validation and test sets are not used due to the limited length of our sample. Moreover, the optimal value for λ is independently selected for each model.

[11] Notice that the last day of the month prior to the day on which re-training occurs—in this case, June 30, 2021—is

are then used to make hourly forecasts for July 2021. To avoid data leakage, the sample means and sample standard deviations of the covariates in Equation (7) are computed using the training set $\mathscr{D}$. This process is repeated for every value in the set Λ.

Using the penalized loss function from Equation (8), an out-of-sample nested model comparison is presented in Table 8. Performance metrics displayed in the table include, for the four versions of the model, the likelihood $\mathscr{L}$, the continuously ranked probability score (CRPS) and the mean absolute error (MAE). The last two are only displayed for the optimal hyperparameter λ^*, whereas the likelihood is also shown for $\lambda = 0$. The MAE is computed as the average absolute point prediction error: $\text{MAE} = \sum_{t=1}^{\tau} \frac{|\hat{y}_t - y_t|}{\tau}$, with $\hat{y}_t$ is the mean of the predicted distribution for y_t. The CRPS metric measures the extent to which predicted distributions are concentrated around observations. It is described in more detail in Appendix A.

Across all models, penalization improves the out-of-sample log-likelihood. However, in contrast to the findings using the full sample, it is unclear whether the inclusion of covariates in the frequency component improves the predictive ability of the mixture model. The *full-freq-full-sev* model is best in terms of the out-of-sample CRPS, which means such covariates help to better concentrate the predictive distributions around future observations. Nevertheless, the *null-freq-full-sev* model has the best out-of-sample performance in terms of likelihood. These mixed findings can be explained by the drift in the behavior of the DART spread distribution across the sample; for instance, positive DART spikes are observed more frequently toward the end of the sample data, entailing non-stationary behavior to which out-of-sample performance is more sensitive.

Table 8: Out-of-sample nested model comparison, using LASSO-type penalty.

Model	Nb. Params.	$\mathscr{L}(\lambda = 0)$	$\mathscr{L}(\lambda^*)$	MAE(λ^*)	CRPS(λ^*)
null-freq-null-sev	8	93,961	93,902	20.96	0.71
full-freq-null-sev	28	91,522	91,503	21.56	0.63
null-freq-full-sev	48	90,259	90,257	20.80	0.62
full-freq-full-sev	68	91,041	90,678	20.89	0.61

Notes: The optimized values of λ, denoted λ^*, for the null-freq-null-sev, full-freq-null-sev, null-freq-full-sev, and full-freq-full-sev models are respectively given by: 1000, 10, 0.1, 100.

$\mathscr{L}$ is the negative log-likelihood, CRPS is the continuously ranked probability score, and MAE is the mean absolute error.

2.5 Comparison to the deep quantile regression approach

Machine learning methods for prediction are receiving extensive attention in the literature. As such it is interesting to compare the performance of the model developed in our work to modern machine-learning based forecasting methods. The focus of this chapter is distributional forecasting. Thus, rather than directly applying existing neural network methods that mainly perform point forecasting, we consider using deep quantile regression as a benchmark to evaluate the distributional forecasts made by the mixture model.

As detailed in Koenker and Hallock (2001), the quantile regression loss function for an individual

not included in the training set $\mathscr{D}$ due to the trading restrictions imposed in Section 2.2.1

observation is given by:

$$\mathscr{L}(\Delta_t|\alpha) = \begin{cases} \alpha\Delta_t, & \text{if } \Delta_t \geq 0, \\ (\alpha-1)\Delta_t, & \text{if } \Delta_t < 0, \end{cases}$$

where α is the quantile level (between 0 and 1) and $\Delta_t = y_t - \widehat{\text{VaR}}(\alpha)_t$, and where $\widehat{\text{VaR}}(\alpha)_t$ is the forecast of the level-α quantile of the predicted distribution for time-t DART spread y_t. Such forecast is obtained as the output of a neural network whose inputs are the covariates. The average loss over the entire dataset is then:

$$\mathscr{L}(y, \widehat{\text{VaR}}(\alpha)) = \frac{1}{\tau}\sum_{t=1}^{\tau} \mathscr{L}(y_t - \widehat{\text{VaR}}(\alpha)_t|\alpha).$$

We implement a feedforward neural network in `R` using the `Keras` package. The input layer takes in the vector of covariates $\mathbf{x}_t$ (previously defined in Equation 4), the hidden layers each contain 10 neurons with ReLU activation functions, and the output layer consists of a single neuron (without an activation function) used to predict the quantile. We tested both one-layer and two-layer neural networks[12]. The Adam optimizer (see Kingma (2014)) with a learning rate of 0.01 is used to minimize the loss function, and the model is trained for 100 epochs.

To compare the quantile forecasts provided by our mixture method against deep quantile regression, the metric considered is the Actual over Expected (AE) VaR exceedance ratio. It is defined as:

$$\text{AE} = \frac{\sum_{t=1}^{\tau} \mathbb{1}_{(y_t > \widehat{\text{VaR}}(\alpha)_t)}}{\tau(1-\alpha)}.$$

The Actual over Expected VaR exceedance ratio consists of the actual number of times the VaR is exceeded in the sample over the expected number of times it is to be exceeded. For example, the VaR(99%) is expected to be exceeded 1% of the time.

AE VaR exceedance ratios are displayed in Table 9 for both the *full-freq-full-sev* version of our model and the deep quantile regression approach (DQR-1 and DQR-2, for the one- and two-layers neural networks, respectively). Such ratios are reported over both the training set (in-sample), and out-of-sample using the sequential validation framework detailed in Section 2.4.3. For the in-sample performance, none of the three strategies dominates the two others. Conversely, out-of-sample, the mixture model is evidently better, except for the VaR(50%), and a near-tie performance with DQR-1 for VaR(5%). The strength of our mixture model lies in being able to adequately capture the tails of the distribution. Despite additional flexibility, the neural network models therefore do not exhibit better predictive performance in our experiment.

2.6 Conclusion

This chapter proposes a method to forecast the distribution of electricity DART spreads. Such a task is non-trivial due to the complex stylized facts of DART spreads, which include the presence of both negative and positive spikes, seasonal behaviors, and complex interactions with multiple dependent covariates. The proposed model is a three-regime mixture model, where spike regimes rely on

[12]Although more than two layers were not tested due to time constraints, we acknowledge that experimenting with additional hidden layers is warranted.

Table 9: Actual over Expected VaR exceedance ratios, *full-freq-full-sev* vs *deep learning*.

Confidence level α	5%	10%	50%	90%	95%
In-sample					
full-freq-full-sev	1.049	1.081	0.965	0.998	1.003
DQR-1	0.842	1.100	1.001	1.022	1.006
DQR-2	1.225	1.054	0.963	1.019	0.998
Out-of-sample					
full-freq-full-sev	1.385	1.199	0.902	0.983	0.996
DQR-1	1.382	1.313	0.972	0.943	0.949
DQR-2	1.822	1.269	1.012	0.937	0.940

Notes: The table entries correspond to the Actual over Expected VaR exceedance ratios at the $\alpha = 5\%$, $\alpha = 10\%$, $\alpha = 50\%$, $\alpha = 90\%$, and $\alpha = 95\%$ VaR confidence levels. The out-of-sample *full-freq-full-sev* model is trained using λ^*.

the Generalized Pareto Distribution, which has the ability to generate extreme values. Regime probabilities and conditional DART spread distributions in each of the regimes are both dependent upon explanatory variables.

The model is implemented on hourly data from the Long Island zone of the NYISO; it is shown to provide an adequate fit to the data. In-sample nested model comparisons highlight the importance of including the entire set of covariates, in both the frequency and severity components of the mixture model, although their inclusion in the severity component is more critical. In addition, the SAGE variable importance assessment algorithm indicates that on a relative basis, the natural gas futures price has the most significant effect upon model performance. The use of regularization methods is explored and is shown to improve the out-of-sample model performance on our data. Finally, a benchmarking experiment reveals that neural-network-based quantile regression benchmarks fail to outperform our mixture model on our test set, highlighting that the better interpretability of our model does not come at the expense of accuracy.

Future areas of investigation include: (1) assessing the model's robustness to alternative data sets, i.e. other zones in the NYISO or other power markets, (2) modeling the joint distribution between multiple zones, which can be used to devise locational differential trading strategies, and (3) using mixture density (neural) networks, see for instance (Bishop, 1994), to obtain additional flexibility to represent the relationship between the DART spread and explanatory variables when producing predictive distributions.

3 Do energy futures add value to stock portfolios?

3.1 Introduction

In light of the major financial crises of the past two decades, commodities have emerged as an attractive asset class for institutional investors, in part due to the diversification benefits they might provide. We therefore investigate whether commodities, more specifically, energy futures, can help to diversify a stock portfolio and thereby improve its risk-adjusted performance.

The literature documenting diversification gains from adding commodities to stock portfolios is extensive. A widely cited work by Geman (2005) reports that unlike stocks, changes in commodity prices are mainly affected by weather, geopolitics, and supply factors. In an influential study by Gorton and Rouwenhorst (2006), the authors highlight that commodities, such as crude oil, have historically had low correlations with stocks and bonds, making them important diversification tools. Moreover, even with 10 additional years of data, Bhardwaj et al. (2015) find that the original conclusions of Gorton and Rouwenhorst (2006) continue to hold, but with an increased correlation observed during periods of crisis. Using the Standard and Poor's Goldman Sachs Commodity Index (GSCI) and the Dow Jones-UBS Commodity Index (DJ-UBSCI), Belousova and Dorfleitner (2012) observe that the inclusion of energy commodities improves portfolio performance in both bull and bear markets. Giamouridis et al. (2014) further suggest that the integration of GSCI commodities using factor-based strategies (e.g. momentum, basis) can also improve portfolio performance. Bessler and Wolff (2015) find that although the out-of-sample diversification benefits are lower than previously suggested by in-sample studies, they are still present. Moreover, they find that portfolio benefits are time-varying, with energy, for example, having a negative impact on the Sharpe ratio between 1995 and 2000 but a positive impact between 2002 and 2009. Drawing on data from the S&P 500 energy sector and crude oil markets, Nguyen et al. (2020) propose a stochastic dominance approach, finding that the inclusion of energy commodities can lead to greater diversification benefits, especially during periods of high economic uncertainty and economic downturns.

Many studies highlight that diversification benefits from energy commodities vary over time. For example, the paper by Jensen et al. (2002) indicates that the addition of GSCI energy commodities during periods of restrictive monetary policy can improve portfolio returns, but leads to poor performance during periods of expansive monetary policy. Du (2005) similarly finds that the benefits obtained are regime dependent, with the greatest advantage offered when equity markets experience high levels of downside risk. Consistent with this perspective, Demidova-Menzel and Heidorn (2007) report that the greatest benefits are derived during inflationary periods. Leveraging the Gorton-Rouwenhorst index, Cheung and Miu (2010) find that commodity futures primarily provide diversification benefits during bull markets, with unconvincing benefits during bear markets. Relying on crude oil and natural gas data[13], Liu and Tu (2012) find that spillover effects between energy futures reduce the potential for diversification during tranquil periods, but not during crisis periods.

Other papers question the ability to use energy commodities as diversifiers, thus challenging the previous conclusions. For example, based on GSCI, DJ-UBSCI, and crude oil data, Daskalaki and Skiadopoulos (2011) do not find significant out-of-sample benefits from incorporating energy commodities. Similarly, the results in Hansen-Tangen and Overaae (2015) do not support the

[13] Liu and Tu (2012) use daily data, whereas previous studies consider a monthly frequency.

conclusion that GSCI energy commodities yield out-of-sample risk-adjusted benefits when added to a portfolio of stocks and bonds. As documented by Elsayed et al. (2020), the correlations between energy markets[14] and global financial markets vary over time, further complicating the ability to use energy commodities as stable diversifiers. By examining crude oil and natural gas data, Lean et al. (2023) demonstrate that an all-equity portfolio stochastically dominates portfolios that include energy futures, using a sample data set covering the period from 1990 to 2017. As is evident from this discussion, the net benefit of using commodities as diversifiers is ambiguous.

The increased financialization of commodity markets further complicates their potential for diversification. As described in Carmona (2015), financialization refers to the increased role of institutional investors in commodity markets, as well as the growth in financially settled contracts in relation to their physically settled counterparts. In the seminal paper by Tang and Xiong (2012), the authors show that financialization has increased the correlation between commodities, contributing to boom and bust cycles. In their foundational study, Silvennoinen and Thorp (2013) report that higher levels in the VIX index[15] are associated with increased commodity-stock correlation. In their frequently cited paper, Creti et al. (2013) emphasize that energy commodities exhibit time-varying correlations with equity markets, with increased financialization occurring after the 2007-2008 financial crisis, potentially influencing their diversification potential. In a notable study, Cheng and Xiong (2014) further document that although financial investors provide liquidity to hedgers to help accommodate their hedging needs, they may also need to quickly unwind their commodity positions (in an attempt to reduce risk) following unexpected price drops in other markets, thereby transmitting outside shocks to commodities. They also find that financialization has made it difficult to differentiate whether movements in futures prices are due to financial trading or actual changes in fundamentals. Using commodity-linked notes (CLN) as a dataset, Henderson et al. (2015) demonstrate that financialization affects commodity futures prices as a result of hedge trades carried out by CLN issuers. In the important paper by Basak and Pavlova (2016), the authors find that commodity futures prices, volatilities, and correlations increase with financialization, more so for index commodities than for non-index commodities. Adams et al. (2020) note that due to financialization, commodities are unlikely to provide effective portfolio diversification benefits, with financial variables explaining a large proportion of the volatility in commodity futures. Baker (2021) highlights that financialization increases the volatility of futures prices and decreases their associated risk premiums. Kang et al. (2023) indicate that financialization significantly increased the correlation between commodity and stock market returns, as well as the pairwise correlation between indexed commodity futures.

The diversification potential of crude oil and natural gas has been extensively explored in the literature. Using 150 years of historical data, Narayan and Gupta (2015) find that oil prices help predict stock returns, with negative oil price changes being more predictive than positive oil price changes. In Christoffersen and Pan (2018), the authors show that after financialization, oil volatility became a strong predictor of stock market returns and volatility. That is, stocks with negative exposure to the option implied volatility of oil earned higher returns compared to those with positive exposures. Moreover, a hedge portfolio based on oil volatility risk outperformed those constructed using risk factors such as momentum, value, and size. Degiannakis et al. (2018) demonstrate that

[14]Such as the World Energy Price Index (WEPI) and crude oil markets. Moreover, Elsayed et al. (2020) use daily data, whereas previous studies consider a monthly frequency.

[15]The Chicago Board of Exchange Volatility Index (VIX index) measures market expectations for the volatility of the S&P 500 index.

higher oil prices tend to be associated with lower stock returns for oil-importing economies and conversely for oil-exporting economies. More precisely, supply-side or precautionary demand oil shocks were associated with negative movements in stock prices, whereas aggregate demand shocks were positive. Oil price volatility was also found to impact stock price volatility, with asymmetric effects (the negative response of stocks to oil price increases was larger in magnitude than the positive response to oil price declines). Gao et al. (2022) conclude that oil volatility increases oil inventory and decreases oil consumption, while causing equity prices to fall, particularly in oil-sensitive industries. In Demiralay et al. (2019), the authors find that commodity futures, such as natural gas, provide diversification benefits that are higher than those provided by stock-only portfolios during periods of global uncertainty, such as the 2008 financial crisis. In Gaete and Herrera (2023), the authors find that energy commodities, particularly natural gas, may offer some potential to reduce portfolio volatility.

Crises and geopolitical conflicts have also been shown to impact the diversification potential of energy commodities. Noguera-Santaella (2016) find that geopolitical events positively impacted oil prices prior to the year 2000, but with negligible impact afterwards. Borgards et al. (2021) establish that the frequency and severity of price overreactions increased during the COVID-19 pandemic, offering the potential for profitable trading opportunities. In their analysis, Adekoya and Oliyide (2021) further reveal that the COVID-19 pandemic caused increased connectedness and risk transmission across commodity and financial markets. Recent research by Zakeri et al. (2022) highlights the impact of the COVID-19 pandemic and the energy crisis triggered by the Russia-Ukraine war on the global energy system, resulting in disruptions in energy supply chains and demand patterns. Chishti et al. (2023) reveal that the Russia-Ukraine war caused a downturn in the crude oil markets, while natural gas experienced significant benefits. Zhang et al. (2024) underscore that the Russia-Ukraine war led to an increase in oil volatility and fundamentally changed the trend of crude oil prices. This analysis therefore reveals the importance of understanding the behavior of energy markets during crisis periods.

In light of this discussion, we therefore reexamine whether incorporating energy futures can add value to an equity portfolio by analyzing an updated data set that includes two new crises, namely the COVID-19 pandemic and the Russia-Ukraine war, which have significantly impacted the energy sector. While previous studies have focused mainly on the increased volatility and changing correlation patterns during these events, our analysis advances the discussion by assessing the financial performance of equity portfolios enhanced with energy commodities. Analyzing their out-of-sample performance during these turbulent periods therefore provides novel insights into their effectiveness under stressed market conditions. In contrast to several studies that omit electricity as a feasible investment opportunity, we additionally include electricity (MISO Texas Hub) alongside crude oil (WTI Cushing) and natural gas (Henry Hub). Electricity is a fundamentally different product from crude oil and natural gas as it cannot be stored on a large scale, thus leading to complex price dynamics and making it difficult to model. In addition, looking ahead, the growing interest in renewable energy generation may potentially increase the appetite to incorporate electricity into portfolio management strategies.

Our model empirically forecasts asset expected returns and volatilities and can adequately handle the non-stationary nature of the time series data. The ability to tackle non-stationary dynamics is crucial in our framework due to important structural breaks occurring during crises. Asset allocation

is performed using a mean-variance optimization approach, with the Sharpe ratio[16] used as the objective function. We benchmark our performance against a static investment in the S&P 500. Our results suggest that despite the increased financialization of commodity markets, energy futures can improve the out-of-sample risk-return performance, resulting in diversification benefits and additional profits relative to the stock-only benchmark portfolio, with the greatest benefits achieved during crisis periods.

The remainder of this chapter is organized as follows. In Section 3.2, we describe the data. In Section 3.3, we describe the asset allocation strategy, the portfolio optimization approach, and the estimation procedure. In Section 3.4, we discuss the out-of-sample portfolio performance in relation to our established benchmark, provide a discussion of the results, and present a sensitivity analysis to assess the robustness of our model. In Section 3.5, we conclude.

3.2 Description of the data

To reflect the actual positions taken by institutional investors, total return data for the S&P 500 (which account for dividends) are used. Futures price data are utilized for energy commodities. The energy commodities considered include electricity, crude oil, and natural gas. The electricity futures data are sourced from the Intercontinental Exchange (ICE), with ticker symbol TDP (MISO Texas Hub), whereas the crude oil and natural gas futures data are obtained from Bloomberg with ticker symbols CL1 (WTI Cushing) and NG1 (Henry Hub), respectively.

The front-month (nearest-to-delivery) futures contracts with a monthly delivery period serve as the basis for the analysis, and futures are rolled over at the close of trading on the last business day that falls on or before the 20th calendar day of the month prior to the contract month. When computing the return on the rollover date t, we ensure that the futures contract at time $t+1$ coincides with the futures contract at time t.

The risk-free rate data are taken from the St-Louis FRED, where we use the market yield on U.S. treasury securities at 1-month constant maturity, quoted on an investment basis per annum.

The data set ranges from May 12, 2014 to May 21, 2024, consisting of 524 weekly observations. The weekly excess returns (returns for the futures) for the various assets are respectively given by

$$E_{t+1}^{(S)} = \frac{S_{t+1} - S_t}{S_t} - r_{t+1}, \quad E_{t+1}^{(i)} = \frac{F_{t+1}^{(i)} - F_t^{(i)}}{F_t^{(i)}},$$

for $i \in \{\text{Electricity (E), Crude oil (C), Natural gas (N)}\}$, where r_{t+1} denotes the weekly risk-free rate prevailing in the time interval $[t, t+1)$. We denote by S_t and $F_t^{(i)}$ the time-t values of the equity asset and i^{th} energy futures, respectively. Unlike the S&P 500 asset, entering futures contracts positions does not require an initial capital outlay[17], which justifies looking at the return rather than the excess return. For clarity, from this point forward, the excess return will refer specifically to the

[16]The Sharpe ratio is defined as $\frac{\mathbb{E}[R^{(X)}-r]}{\sigma^{(X)}}$, where $R^{(X)}$ is the return for asset X, r is the risk-free rate, and $\sigma^{(X)}$ is the volatility for asset X. The Sharpe ratio is essentially the inverse of the coefficient of variation, adjusted for the risk-free rate.

[17]We acknowledge that futures positions do require margin capital in practice (which may allow for treating positions on futures on a more comparable basis to investments in the S&P 500), hence this is a simplifying assumption that we make.

return for the futures assets. Excess returns can thus be understood as the excess return over the cost of carry associated with the investment instrument.

Table 10: Descriptive statistics of weekly excess returns.

	Min	Q1	Median	Q3	Max	Mean	Std	Skew	Exc. Kurt
S&P 500	-0.150	-0.008	0.004	0.015	0.121	0.002	0.023	-0.644	6.667
Electricity	-0.636	-0.031	-0.003	0.029	1.714	0.003	0.139	5.011	58.817
Crude oil	-0.293	-0.027	0.003	0.031	0.464	0.002	0.060	0.953	10.124
Natural gas	-0.246	-0.047	-0.002	0.041	0.697	-0.002	0.078	1.272	12.243

Notes: The data consists of weekly excess returns, ranging between May 12, 2014 to May 21, 2024, spanning approximately 120 months.

In Table 10, we provide descriptive statistics for asset excess returns. Looking at the standard deviations, we observe that the energy commodities are more volatile than the S&P 500. Moreover, of the assets considered, electricity has the largest kurtosis. Electricity cannot be efficiently stored at a large scale; hence, its supply and demand must always be in equilibrium. Transmission maintenance, unexpected interruptions, and unforeseen demand can therefore give rise to price spikes. Such spikes explain the high kurtosis for electricity futures. We also highlight that each of the energy commodities has a positive skewness whereas the S&P 500 has a negative skewness.

Table 11: Annualized sample Sharpe ratios by period.

	Full	Normal	COVID-19	Russia-Ukraine
S&P 500	0.76	0.76	1.08	0.46
Electricity	0.18	-0.68	0.97	0.48
Crude oil	0.18	-0.50	1.35	-0.18
Natural gas	-0.18	-0.53	0.77	-0.56

Notes: The data consists of weekly excess returns, ranging between May 12, 2014 to May 21, 2024. The normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024. The annualized sample Sharpe ratio is computed for each of these periods.

To assess the risk-adjusted performance across the different assets while tracking their fluctuations over time, we present the sample annualized Sharpe ratios for different periods of the sample in Table 11. Such sample annualized Sharpe ratios are computed as

$$\sqrt{52}\frac{\bar{E}_t^{(X)}}{\hat{\sigma}_t^{(X)}},$$

where the sample mean and sample standard deviation of the weekly excess returns for asset $X \in \{S, F^{(E)}, F^{(C)}, F^{(N)}\}$ over the corresponding period, as estimated from n observations, are

given by

$$\bar{E}_t^{(X)} = \frac{1}{n}\sum_{u=0}^{n-1} E_{t-u}^{(X)},$$

$$\hat{\sigma}_t^{(X)} = \sqrt{\frac{1}{n-1}\sum_{u=0}^{n-1}\left(E_{t-u}^{(X)} - \bar{E}_t^{(X)}\right)^2}.$$

The full sample is broken down into the following three periods. The normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024.

The S&P 500 has a positive Sharpe ratio over the full sample period as well as over each sub-period. For energy futures, we focus on the absolute value of the Sharpe ratios because we can take either long or short positions. The energy futures have smaller Sharpe ratios, in absolute value, during the normal period. During the COVID-19 period, crude oil has the largest Sharpe ratio, while electricity and natural gas have lower Sharpe ratios than the S&P 500. During the Russia-Ukraine period, the opposite occurs: both natural gas and electricity yield higher absolute Sharpe ratios relative to the S&P 500, whereas crude oil underperforms the S&P 500. This suggests that adding energy futures to an equity portfolio has the potential to improve its risk-adjusted profile under certain market conditions.

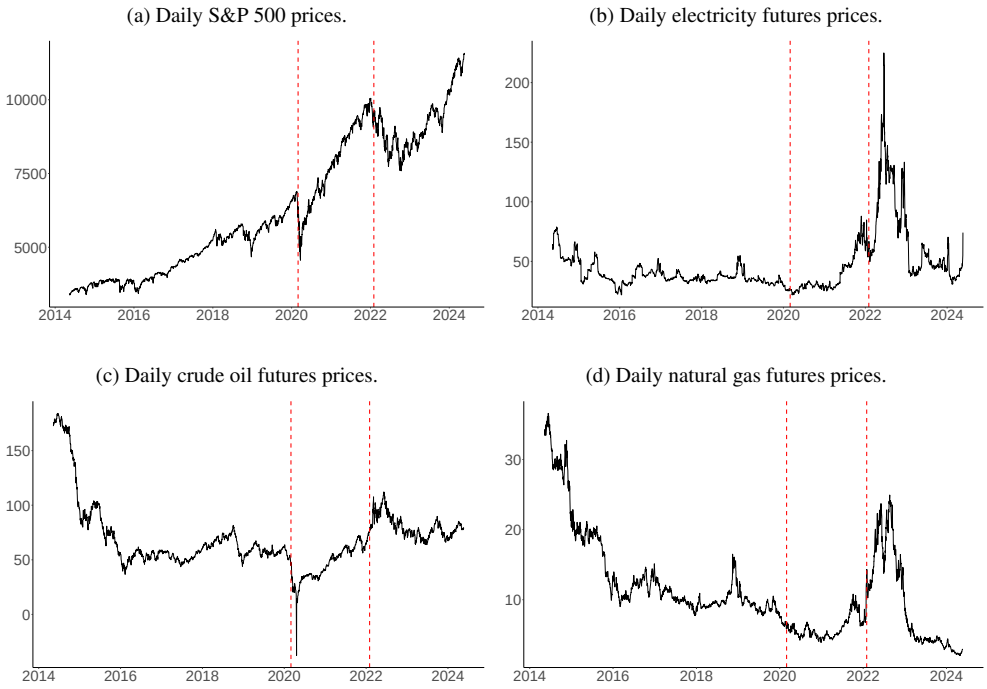
It is also interesting to note the signs of the Sharpe ratios across the various sub-periods. For instance, during the normal period, the Sharpe ratios for the energy futures are negative. That is, under normal conditions, investors are willing to pay a premium to secure the physical commodity. Over the COVID-19 pandemic, the Sharpe ratios turn positive. At the start of the pandemic, the energy sector was depressed due to travel restrictions and economic lock-downs, resulting in excess supply and even negative oil prices. By the end of the pandemic, however, there was renewed economic activity and pent-up demand.

In Figure 8, we display the price time series for each of the assets. At the beginning of the COVID-19 period (i.e., the first vertical red bar), prices were low; but by the end of the pandemic (i.e., the second vertical red bar), prices rose sharply, thus resulting in an overall net positive performance over this period. Then, during the Russia-Ukraine period, the Sharpe ratios once again turn negative (except for electricity). This can be accounted for by the increased uncertainty surrounding the energy sector during this time-frame, characterized by supply chain bottlenecks and associated supply shortages.

Figure 9 displays the time series of weekly excess returns for the various assets. Panels (a) and (c) respectively indicate that the S&P 500 and crude oil react more extensively to the COVID-19 pandemic, as indicated by the increased variability in the associated excess returns. Alternatively, in Panels (b) and (d), electricity and natural gas respond more intensely to the Russia-Ukraine period, once again evidenced by the corresponding rise in variability of the related excess returns. Although electricity is more volatile than the other assets during the Russia-Ukraine period, it can still potentially help enhancing risk-adjusted performance, as evidenced by its relatively large Sharpe ratio during that period (see Table 11).

In Table 12, we present the correlations between the excess returns across the various periods considered. In each panel, the entries above the diagonal correspond to Spearman correlations and the entries below the diagonal to Pearson correlations. We include Spearman correlations since they

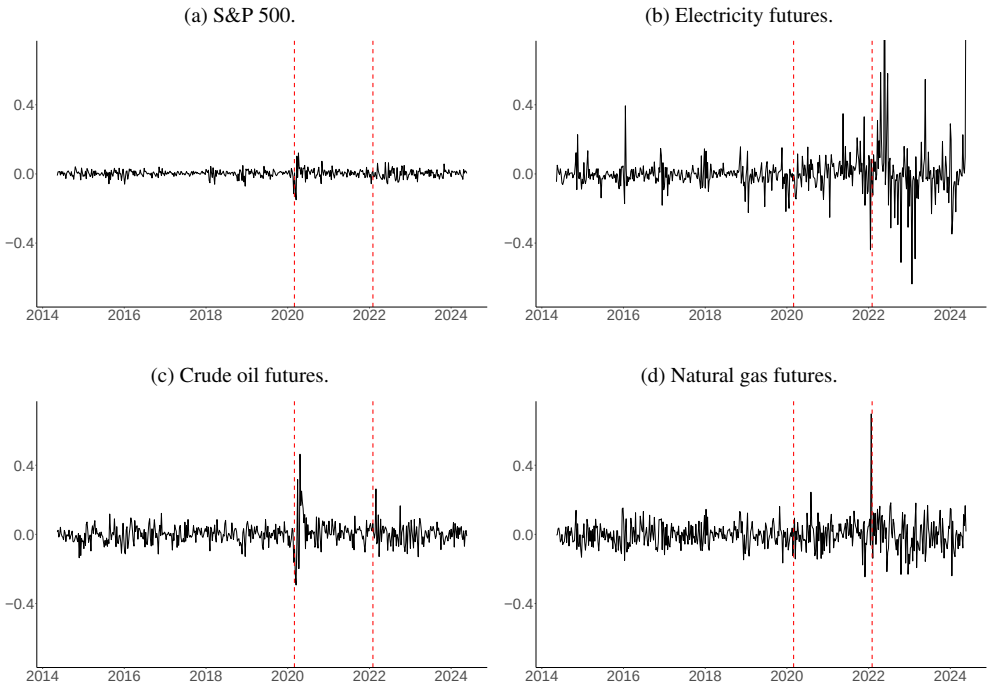
Figure 8: Time series of daily asset prices.



Notes: The data consists of daily prices, ranging between May 12, 2014 to May 21, 2024. The vertical bars demarcate the different periods considered. That is, the normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period (which includes the market down-turn) ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024.

are not influenced by outliers, unlike Pearson correlations. The Pearson and Spearman correlation values are quite similar and of the same order of magnitude, with the exception of electricity and natural gas, for which the Spearman correlation is roughly twice the Pearson correlation in the full and Russia-Ukraine periods. This suggests that the relationship between electricity and natural gas is non-linear and characterized by spikes.

Figure 9: Time series of weekly excess returns.



Notes: The data consists of weekly excess returns, ranging between May 12, 2014 to May 21, 2024. The vertical bars demarcate the different periods considered. That is, the normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024. In Panel (b), the spikes on May 23, 2022 and May 20, 2024 reach roughly 1.7 and 1.3, respectively, far exceeding the y-axis limit of 0.7 imposed for clarity.

Table 12: Pearson correlations of weekly excess returns, by period.

	Full				Normal			
	S&P 500	Electricity	Crude oil	Natural gas	S&P 500	Electricity	Crude oil	Natural gas
S&P 500	1	**0.05**	**0.24**	**0.10**	1	**0.10**	**0.26**	**0.10**
Electricity	0.10	1	**0.11**	**0.59**	0.11	1	**0.13**	**0.65**
Crude oil	0.28	0.04	1	**0.14**	0.41	0.13	1	**0.14**
Natural gas	0.13	0.28	0.11	1	0.13	0.59	0.15	1
	COVID-19				Russia-Ukraine			
	S&P 500	Electricity	Crude oil	Natural gas	S&P 500	Electricity	Crude oil	Natural gas
S&P 500	1	**0.05**	**0.31**	**0.19**	1	**-0.02**	**0.12**	**0.03**
Electricity	0.11	1	**0.03**	**0.62**	0.12	1	**0.09**	**0.43**
Crude oil	0.24	-0.05	1	**0.08**	0.10	0.05	1	**0.14**
Natural gas	0.18	0.42	0.04	1	0.04	0.19	0.15	1

Notes: The data consists of weekly excess returns, ranging between May 12, 2014 to May 21, 2024. The normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024. Pearson (below the diagonals) and Spearman (bold numbers above the diagonals) correlations are presented.

In general, the correlations between the assets are weak. An exception is electricity and natural gas, which have a higher correlation relative to the other pairs of assets. Natural gas is an important

fuel in electricity generation and often serves as the marginal unit of production to quickly balance supply and demand in real time, thus making it a significant driver of electricity prices. Nevertheless, during the Russia-Ukraine war, their correlation weakens, which helps reduce portfolio volatility by means of diversification. Moreover, in that same time period, the signs of their Sharpe ratios decouple, further highlighting their diversification potential.

In the top right, bottom left and bottom right panels of Table 12, we present the Pearson correlations over the normal, COVID-19, and Russia-Ukraine periods. As previously indicated in the literature, we can observe that the correlations are time-varying. For example, for the S&P 500 and crude oil as well as electricity and natural gas pairs, the correlations are the highest in the normal period, and then they progressively decrease during the two crisis periods.

Overall, our analysis of the descriptive statistics shows that energy commodities can improve the risk-adjusted expected portfolio return, as they demonstrate higher Sharpe ratios compared to the S&P 500 over certain periods. Additionally, they offer diversification benefits, given their generally weak correlations with the S&P 500. However, achieving enhanced investment strategies requires dynamically re-adjusting the portfolio to capture the time-varying characteristics of the various assets, such as the signs and magnitudes of their expected excess returns, their volatilities and corresponding Sharpe ratios, and their correlations.

3.3 Asset allocation

This section outlines the portfolio optimization framework, detailing the allocation strategy across the portfolio's various assets. It also describes the model used to estimate the moments of the asset returns, which are essential to calculate the predicted Sharpe ratios of the different allocation options.

3.3.1 Portfolio optimization

We consider the perspective of an investor seeking to enhance the risk-adjusted returns of an equity portfolio through an overlay consisting of energy futures. The returns for asset X are thus given by

$$R_{t+1}^{(X)} = \frac{X_{t+1} - X_t}{X_t},$$

and the one-week-ahead conditional moments for assets $X, Y \in \{S, F^{(E)}, F^{(C)}, F^{(N)}\}$ are

$$\mu_t^{(X)} = \mathbb{E}_t \left[R_{t+1}^{(X)} \right], \quad \Sigma_t^{(X,Y)} = \mathrm{Cov}_t [R_{t+1}^{(X)}, R_{t+1}^{(Y)}].$$

Since taking positions in futures does not require bearing the cost of capital, our methodology fully allocates the capital to the S&P 500 asset on each time step. As such, the number of S&P 500 shares held at any time t until $t+1$ is W_t/S_t, where W_t is the wealth at time t. Without loss of generality, consider an initial wealth of $W_0 = 1$.

In general, denoting by $N_{F^{(i)}}(t)$ the number of long positions on the i^{th} futures between t and

$t+1$, the wealth at time $t+1$ evolves according to

$$\begin{aligned} W_{t+1} &= W_t \frac{S_{t+1}}{S_t} + \sum_{i \in \{E,C,N\}} N_{F^{(i)}}(t)(F_{t+1}^{(i)} - F_t^{(i)}) \\ &= W_t \left(\frac{S_{t+1}}{S_t} + \sum_{i \in \{E,C,N\}} \frac{N_{F^{(i)}}(t) F_t^{(i)}}{W_t} \frac{F_{t+1}^{(i)} - F_t^{(i)}}{F_t^{(i)}} \right), \end{aligned}$$

which implies that

$$\frac{W_{t+1} - W_t}{W_t} = \frac{S_{t+1} - S_t}{S_t} + \sum_{i \in \{E,C,N\}} \omega_{F_t^{(i)}} \frac{F_{t+1}^{(i)} - F_t^{(i)}}{F_t^{(i)}},$$

where $\omega_{F_t^{(i)}}$ is the weight associated with futures asset i at time t, which can be either positive or negative depending on whether a long or short futures position is used. The weight associated with a futures contract is thus interpreted as the exposure to the corresponding futures position, as a proportion of the portfolio's total nominal wealth. The one-week-ahead conditional mean of the portfolio returns is

$$\mathbb{E}_t \left[\frac{W_{t+1}}{W_t} - 1 \right] = \boldsymbol{\omega}_t \boldsymbol{\mu}_t^{\mathsf{T}}, \tag{9}$$

with $\boldsymbol{\mu}_t = [\mu_t^{(S)}, \mu_t^{(E)}, \mu_t^{(C)}, \mu_t^{(N)}]$ and where $\boldsymbol{\omega}_t = [1, \omega_{F_t^{(E)}}, \omega_{F_t^{(C)}}, \omega_{F_t^{(N)}}]$. The one-week-ahead conditional variance of the portfolio returns is

$$\mathbb{V}_t \left[\frac{W_{t+1}}{W_t} - 1 \right] = \boldsymbol{\omega}_t \boldsymbol{\Sigma}_t \boldsymbol{\omega}_t^{\mathsf{T}}, \tag{10}$$

where $\boldsymbol{\Sigma}_t = [\Sigma_t^{(X,Y)}]$ denotes the conditional covariance matrix of returns. The one-week-ahead Sharpe ratio is

$$\mathrm{SR}_t = \frac{\boldsymbol{\omega}_t \boldsymbol{\mu}_t^{\mathsf{T}} - r_{t+1}}{\sqrt{\boldsymbol{\omega}_t \boldsymbol{\Sigma}_t \boldsymbol{\omega}_t^{\mathsf{T}}}}.$$

The one-week-ahead Sharpe ratio serves as the objective function for portfolio optimization, with the allocation being re-optimized each week based on updated market conditions. Thus, we numerically solve[18] for weights that maximize the Sharpe ratio on a weekly basis. To avoid excessive exposures to any of the energy futures, the weights are constrained to lie in the interval $-K < \omega_{F_t^{(i)}} < K$, for $i \in \{E,C,N\}$, and for some deterministic bound[19] K. The methodology used to estimate the moments in Equations (9) and (10) is discussed in the next section.

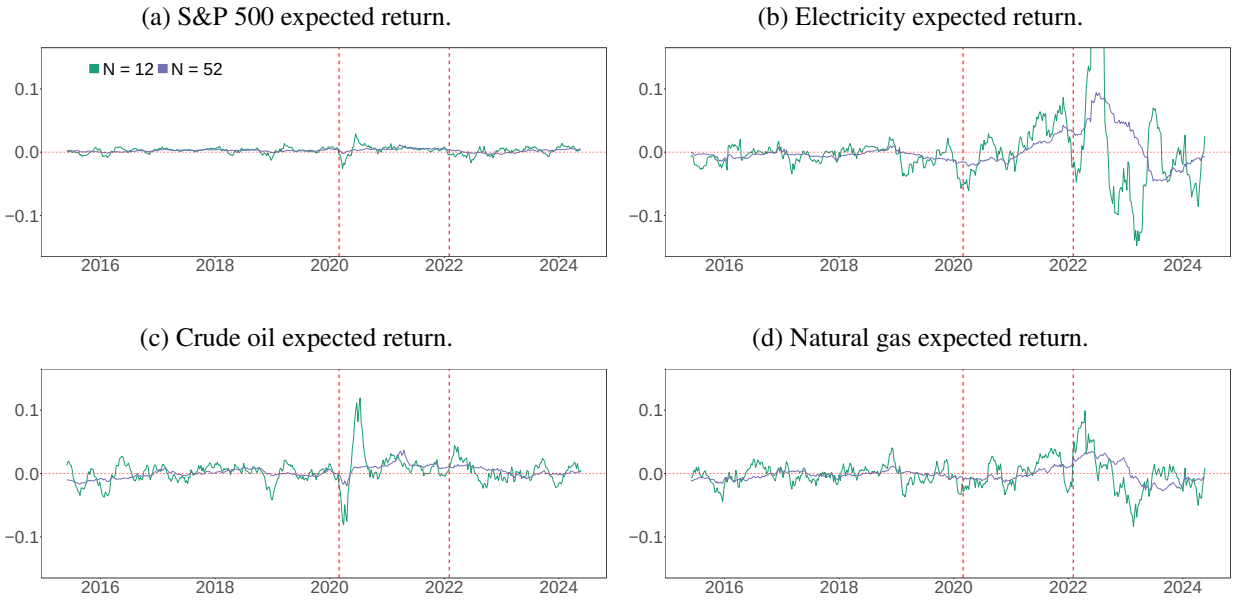
[18] The `solnp` package in `R` is used to perform the optimization numerically.

[19] In Bessler and Wolff (2015), the strategic weights for commodities are fixed (i.e., static) at $K = 5\%$ and $K = 15\%$ for the conservative and aggressive investor clientele, respectively.

3.3.2 Return moments estimation

At any time point t, the calculation of the one-week-ahead Sharpe ratio for a given strategy requires estimates of the conditional means $\boldsymbol{\mu}_t$ and covariances $\boldsymbol{\Sigma}_t$. The approach used to derive such estimates takes into consideration the highly non-stationary nature of our data sample and is thus characterized by changing asset return dynamics. Several conventional econometric models were contemplated for the estimation of volatilities and correlations, such as the Dynamic Conditional Correlation model, GARCH models and Hidden Markov Models. Such models gave unsatisfactory results in unreported tests (see Appendix B); empirical models constructed using a rolling window approach are better able to cope with non-stationarity and can more easily track the time-varying distribution of the underlying data. A sample-based rolling window estimation approach is therefore employed, as outlined in Bessler and Wolff (2015).

Figure 10: One-week-ahead predictions of asset expected returns.



The data is weekly from June 1, 2015, to May 21, 2024, with a rolling window of 12 or 52 weeks. Vertical bars mark three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024). In Panel (b), values between May 30 and August 15, 2022, range from 0.19 to 0.32, exceeding the y-axis limit.

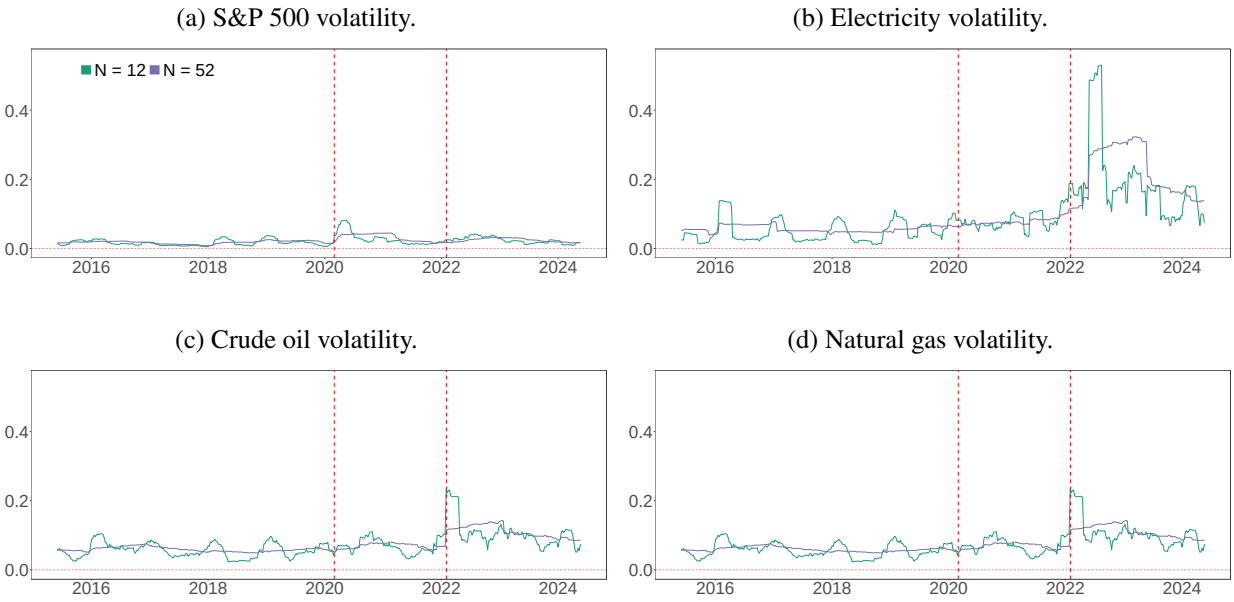
Estimates for the one-week-ahead expected return for asset j and return covariance between assets j and k are

$$\widehat{\mu}_t^{(j)} = \frac{1}{N}\sum_{u=0}^{N-1} R_{t-u}^{(j)}, \quad \widehat{\sigma}_t^{(k,j)} = \frac{1}{N-1}\sum_{u=0}^{N-1}\left(R_{t-u}^{(k)} - \hat{\mu}_t^{(k)}\right)\left(R_{t-u}^{(j)} - \hat{\mu}_t^{(j)}\right),$$

for which the rolling sample means are computed using the last N time steps (i.e., the window size is N periods). This process is repeated at each time point t, where the portfolio is rebalanced.

In Figure 10, the predicted expected asset returns for the $N = 12$ and $N = 52$-week rolling

Figure 11: One-week-ahead predictions of volatilities.



The weekly data spans June 1, 2015 – May 21, 2024, with a rolling window of 12 or 52 weeks. Vertical bars mark three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

window horizons are displayed. For each of the assets considered, the expected returns react negatively after the onset of the COVID-19 pandemic, but then rebound shortly afterwards. In addition, the expected returns for electricity and natural gas become increasingly more volatile in response to the Russia-Ukraine period, whereas the expected returns for the S&P 500 and crude oil remain relatively more stable.

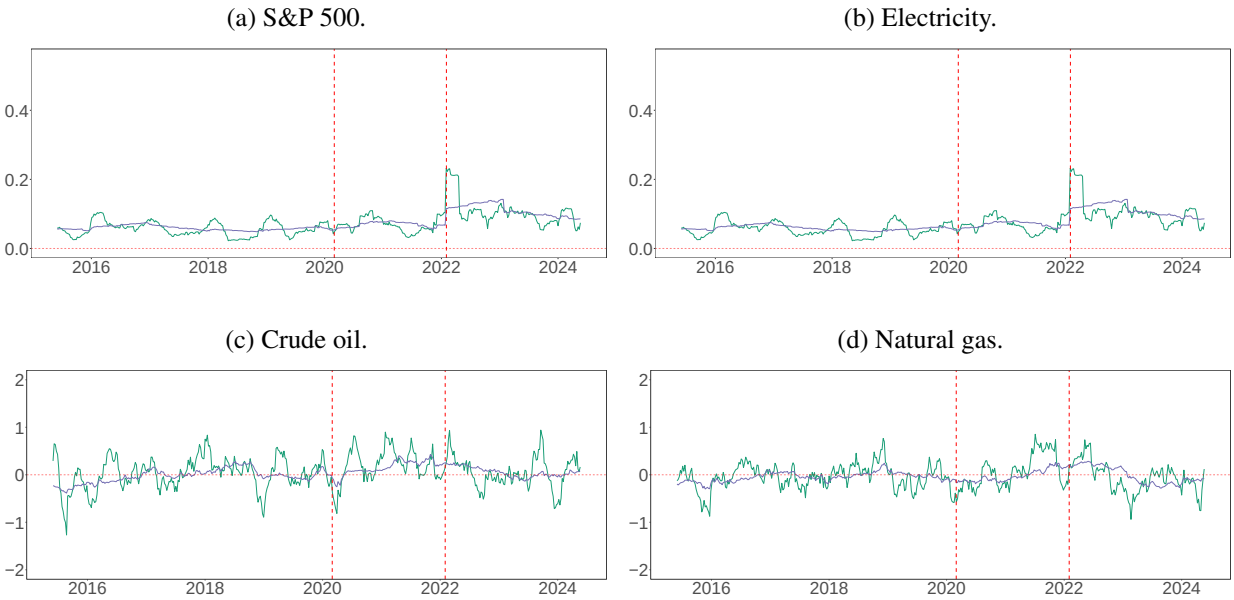
Figure 11 presents the predicted volatilities (i.e., the square root of the predicted variances) for each of the assets at the $N = 12$ and $N = 52$-week rolling window horizons. The volatilities for the S&P 500 and crude oil jump upward in response to the COVID-19 pandemic and then progressively begin to revert to more normal levels. On the other hand, the volatilities for electricity and natural gas jump upward in response to the energy crisis, but do not react as strongly to the COVID-19 pandemic.

In Figure 12, we present the predicted Sharpe ratios for each asset for both the $N = 12$ and $N = 52$-week rolling window horizons.

In Figure 13, we present the predicted asset correlations for the N-week rolling window horizons, with $N \in \{12, 52\}$. Starting with $N = 12$, we can observe that for each of the asset pairs, the correlations oscillate between positive and negative values. An exception is the electricity and natural gas pair for which the correlation is generally positive prior to the energy crisis, but then begins to oscillate between positive and negative values thereafter. However, for $N = 52$, the correlations are generally positive for each of the asset pairs considered, which therefore suggests fewer estimated diversification benefits for larger window sizes.

In Figure 14, we present the optimized portfolio weights time series for each of the energy

Figure 12: One-week-ahead Sharpe predictions.

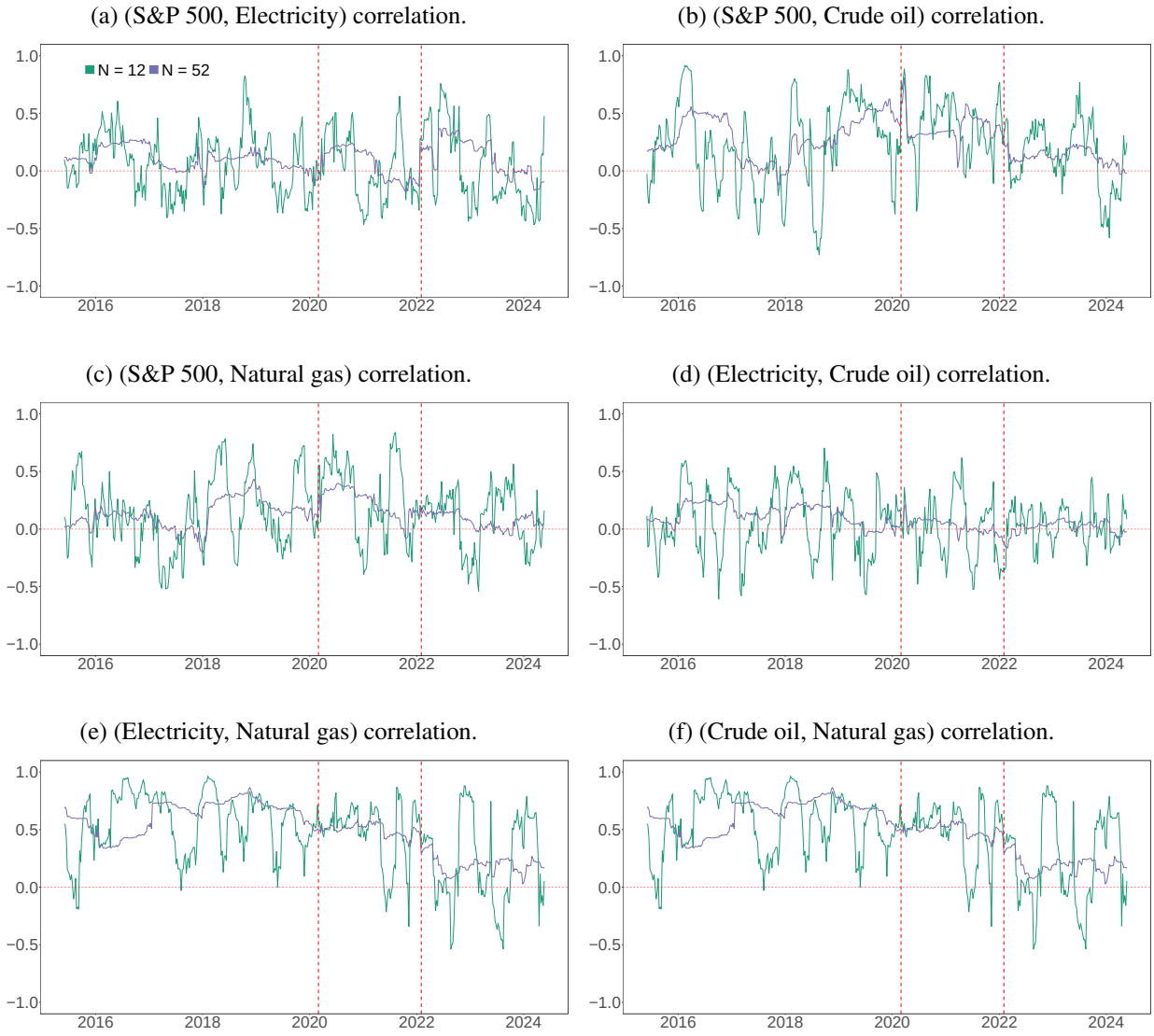


The weekly data spans June 1, 2015 – May 21, 2024, with a rolling window of 12 or 52 weeks. Vertical bars mark three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

futures as a function of both the window size and the imposed bounds. We can observe that the constraints are often binding. This suggests that the bounds are preventing our model from taking on excessively large positions, thus acting as a layer of security for investors[20]. In addition, we can observe that for a given limit K, increasing the window size tends to decrease the variability of the associated futures positions for each of the energy commodities, and therefore leads to more stability and less turnover. However, larger window sizes are associated with slower reactions to crisis events, potentially leading to under-reaction.

[20]Alternative objective functions to the Sharpe ratio that more directly capture the tail risk related to the energy futures positions may also be considered, such as the Sortino ratio.

Figure 13: One-week-ahead correlation predictions.



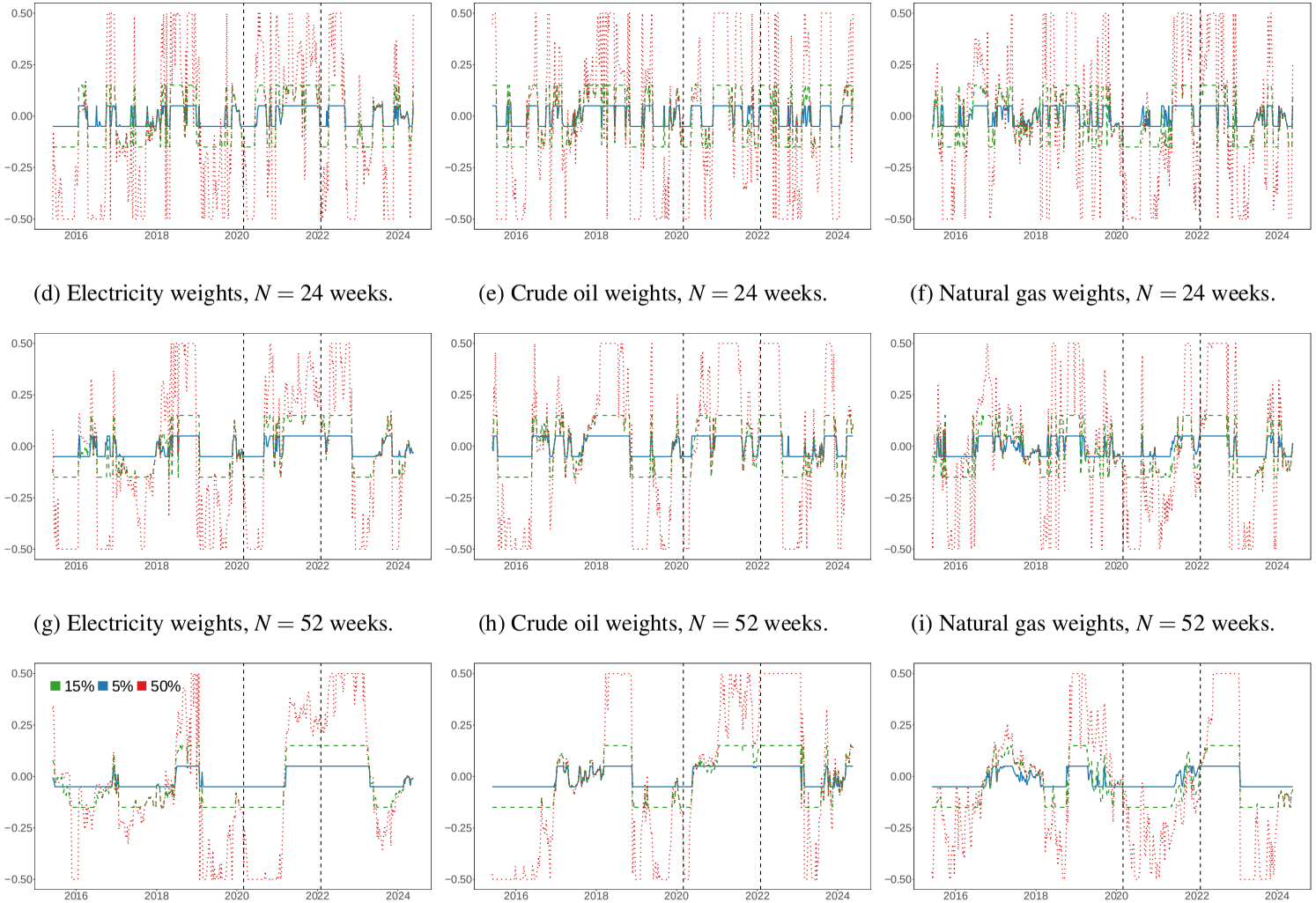
The data is weekly from June 1, 2015, to May 21, 2024. The rolling window spans $N = 12$ or $N = 52$ weeks. Vertical bars mark three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

Figure 14: Portfolio weights for various window sizes and bounds.

(a) Electricity weights, $N = 12$ weeks. (b) Crude oil weights, $N = 12$ weeks. (c) Natural gas weights, $N = 12$ weeks.

(d) Electricity weights, $N = 24$ weeks. (e) Crude oil weights, $N = 24$ weeks. (f) Natural gas weights, $N = 24$ weeks.

(g) Electricity weights, $N = 52$ weeks. (h) Crude oil weights, $N = 52$ weeks. (i) Natural gas weights, $N = 52$ weeks.



Notes: The data granularity is weekly, ranging from June 1, 2015 to May 21, 2024. We measure the sensitivity of the portfolio weights to (1) the rolling window size N and (2) the imposed bounds K (blue curves: 5%, green curves: 15%, red curves: 50%). The rolling window uses either a $N = 12$, $N = 24$, or $N = 52$-week period. The vertical bars demarcate the different periods considered. That is, the normal period ranges between June 1, 2015 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024.

3.4 Performance assessment

3.4.1 Out-of-sample

Performance is evaluated using a back-test, where the out-of-sample portfolio performance is assessed on the period from June 1, 2015 to May 21, 2024. Data prior to June 1, 2015 is excluded from the back-test to ensure a sufficient sample for computing the initial rolling window estimate of the moments.

The results are assessed in terms of the annualized sample Sharpe and Sortino ratios, the maximum drawdown, VaR(1%), and CVaR(1%). The annualized portfolio sample Sharpe ratio is given by

$$\sqrt{52}\frac{\hat{\mu}^{(W)}}{\sqrt{\frac{1}{N-1}\sum_{t=1}^{N}\left(R_t^{(W)}-r_t-\hat{\mu}^{(W)}\right)^2}},$$

where $(R_t^{(W)}-r_t)$ denotes the portfolio excess return at time t and $\hat{\mu}^{(W)}=\frac{1}{N}\sum_{t=1}^{N}(R_t^{(W)}-r_t)$ is the sample mean of portfolio excess returns computed out-of-sample from $t=1,\ldots,N$. The annualized portfolio sample Sortino ratio is defined as

$$\sqrt{52}\frac{\hat{\mu}^{(W)}}{\sqrt{\frac{1}{N-1}\sum_{t=1}^{N}\mathbb{1}_{(R_t^{(W)}-r_t)<0}\left(R_t^{(W)}-r_t\right)^2}},$$

such that $\mathbb{1}_A$ is an indicator variable equal to 1 if event A occurs and 0 otherwise. Finally, the maximum drawdown for the portfolio is calculated as

$$\max_{t\in\{1,\ldots,N\}}\left(\frac{\mathrm{HWM}_t-\mathrm{W}_t}{\mathrm{HWM}_t}\right),$$

where the high-water mark is $\mathrm{HWM}_t=\max_{j\in\{1,\ldots,t\}}W_j$. The maximum drawdown computes the maximum observed loss from a peak to a trough over a time period.

3.4.2 Benchmarks

The performance of the optimized portfolio is evaluated against the following benchmark portfolios, which are static investments in either (1) the S&P 500, (2) electricity, (3) crude oil, or (4) natural gas. Without loss of generality, consider an initial wealth of $W_0^{(S)}=W_0^{(i)}=1$. The wealth at time $t+1$ for the S&P 500 static benchmark portfolio, denoted as $W_{t+1}^{(S)}$, evolves according to

$$W_{t+1}^{(S)}=W_t^{(S)}\frac{S_{t+1}}{S_t},$$

whereas the wealth at time $t+1$ for the static i^{th} energy futures benchmark portfolio, indicated as $W_{t+1}^{(i)}$, is governed by

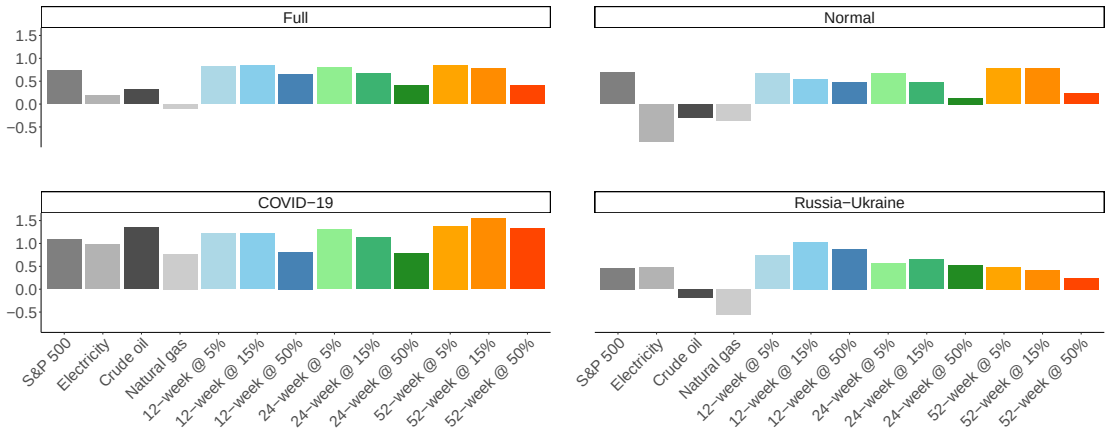
$$W_{t+1}^{(i)}=W_t^{(i)}(1+r_{t+1})+W_t^{(i)}\frac{F_{t+1}^{(i)}-F_t^{(i)}}{F_t^{(i)}}.$$

As previously explained, since futures contracts do not require initial capital, the capital is therefore held in a risk-free asset.

3.4.3 Results

In Figures 15 through 19, we present out-of-sample portfolio performance metrics for each of the optimized portfolios, as well as for the benchmark portfolios. Performance metrics considered include the annualized Sharpe and Sortino ratios, the weekly VaR(1%) and CVaR(1%), and the maximum drawdown. We perform sensitivity tests with respect to the size of the rolling window (i.e., using $N \in \{12, 24, 52\}$ week window sizes) and to the limits imposed on the futures weights during optimization (i.e., at $K \in \{5\%, 15\%, 50\%\}$ of nominal wealth).

Figure 15: Annualized Sharpe ratio.



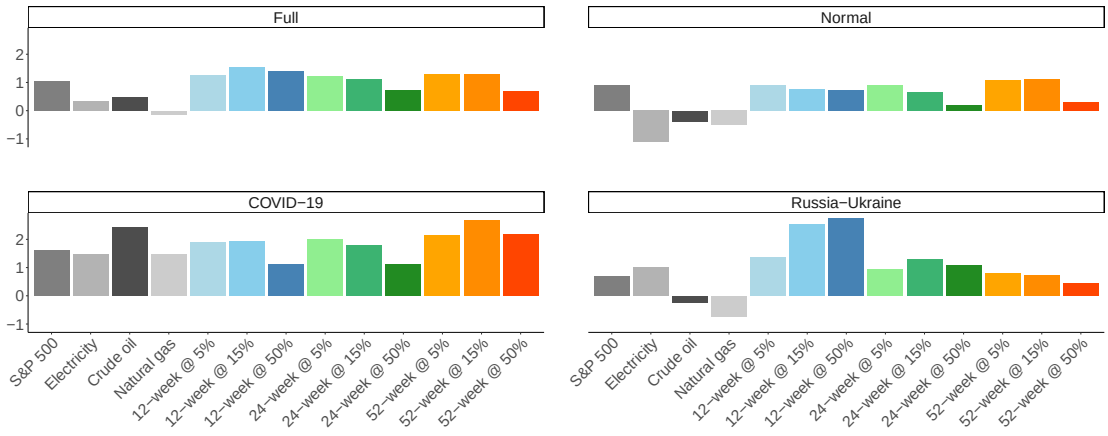
Three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

In Figures 15 and 16, we display the annualized Sharpe and Sortino ratios for each portfolio, across every period. We first look at the benchmark portfolios, which reflect the performance of static positions in individual assets. Amongst these, the S&P 500 has the highest Sharpe and Sortino ratios in absolute value over the full sample. Recall that short positions on futures can be used, which justifies looking at the absolute value. Hence, a static investor selecting a single asset class would be better served with the S&P 500.

However, the performance of the various assets fluctuates substantially across the different periods. For example, the asset with the highest Sharpe and Sortino ratios in absolute value is electricity during the normal period, crude oil during the COVID-19 period, and natural gas during the Russia-Ukraine period (for the Sharpe ratio only). This highlights the importance of using a dynamic portfolio strategy that adapts to fluctuating prospective return distributions, which may thus yield superior risk-adjusted expected returns.

Both the position limit K and the window size N impact the performance of dynamic portfolios. Regardless of the window size N, adopting more conservative bounds (i.e., $K \in \{5\%, 15\%\}$) generally enhances the Sharpe and Sortino ratios compared to more aggressive bounds (i.e., $K = 50\%$). Indeed, in almost all periods and for most window sizes N, the Sharpe ratio and Sortino

Figure 16: Annualized Sortino ratio.



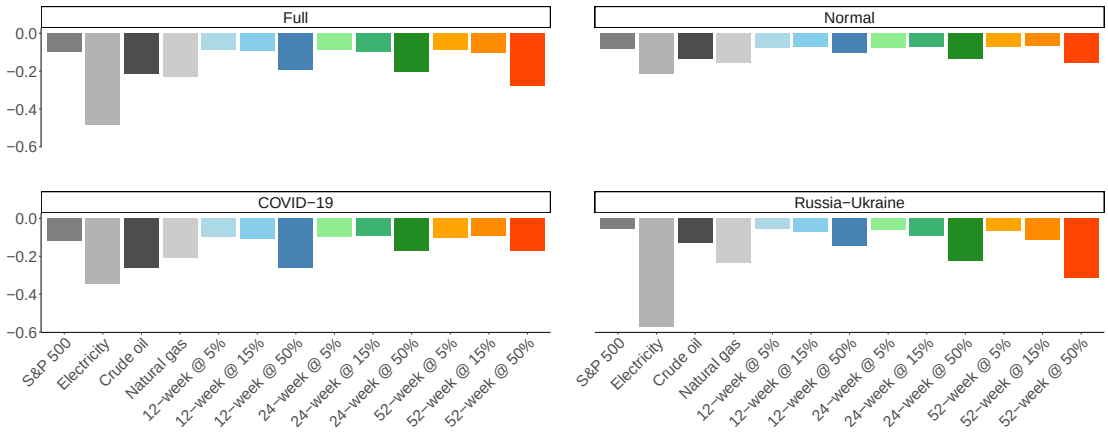
Three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

ratios obtained using $K = 50\%$ are amongst the smallest. Furthermore, as highlighted in Figures 17, 18, and 19, the corresponding VaR(1%), CVaR(1%), and maximum drawdown also show higher levels of risk when using $K = 50\%$. We therefore choose to focus our remaining analysis on $K \in \{5\%, 15\%\}$. In either case, using $K = 5\%$ or $K = 15\%$, the VaR(1%), CVaR(1%), and maximum drawdown values in Figures 17, 18, and 19 indicate comparable tail risk.

With respect to the window size N, when market conditions suddenly change, shorter window sizes can perform better by adjusting more quickly. For example, when transitioning from the COVID-19 period to the Russia-Ukraine period, the corresponding Sharpe and Sortino ratios for the $N = 12$ week portfolios are the highest; this can be seen by looking at the blue bars on the bottom right sub-panels of both Figure 15 and Figure 16.

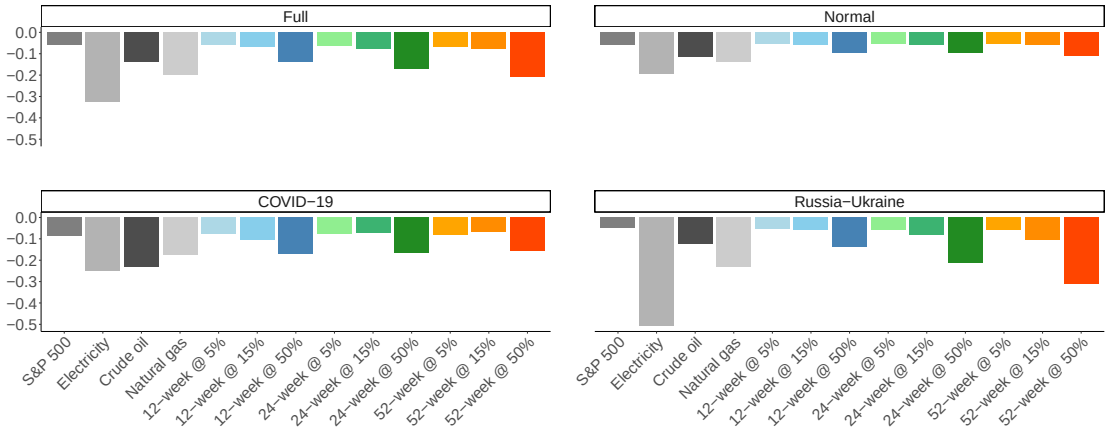
In Figures 17, 18, and 19 we show the VaR(1%), CVaR(1%), and maximum drawdown for the various strategies. For any window size N, portfolios using $K \in \{5\%, 15\%\}$ exhibit VaR(1%), CVaR(1%), and maximum drawdown metrics that are comparable to those of the S&P 500 benchmark across every period considered. This suggests that adding energy commodities does not inherently increase the tail risk of the portfolio due to the diversification benefits achieved.

Figure 18: CVaR(1%).



Three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

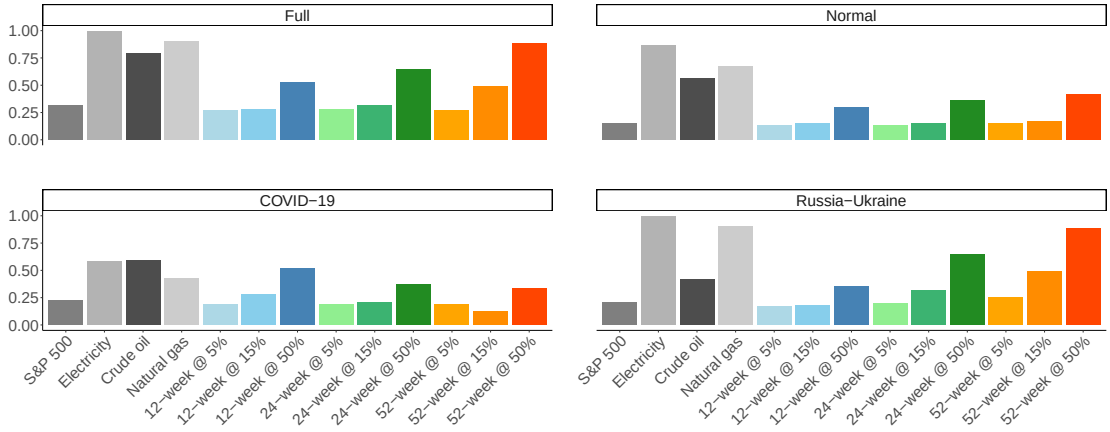
Figure 17: VaR(1%).



Three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

The portfolio wealth curves are displayed in Figure 20, which indicate how a dollar that is initially invested in a portfolio grows over time. Relative to the S&P 500, the $K \in \{5\%, 15\%\}$ portfolios consistently achieve higher relative wealth, thus indicating the robustness of our model in relation to the window size used to estimate asset return moments. Moreover, in Panels (b), (c), and (d), we can observe that the wealth curves are close together during the normal period, and only begin to differentiate substantially during the COVID-19 crisis (onward). This implies that most of the benefits of including energy futures are realized during the crisis periods, which is aligned with the existing literature, which often finds that the greatest diversification benefits from energy commodities occurs during turbulent periods. This finding, however, is retrospective. That is, using our methodology, portfolio managers can only recognize a regime change after the fact, thereby introducing a delay between the start of the new regime and its detection by the traders.

Figure 19: Maximum drawdown.

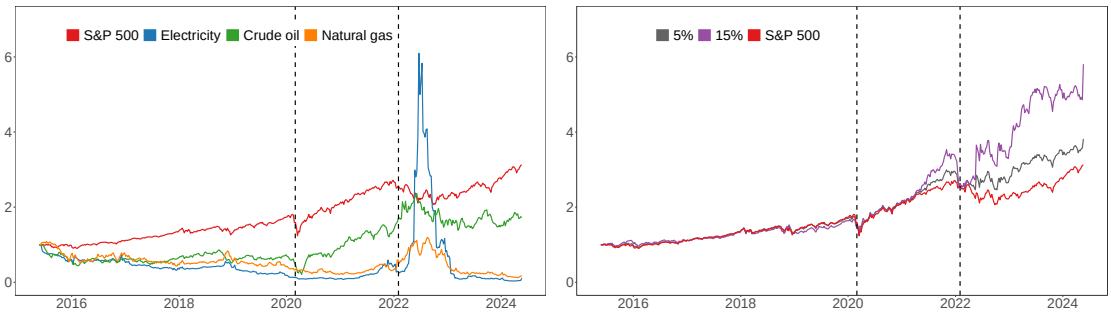


Three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

Figure 20: Portfolio wealth curves.

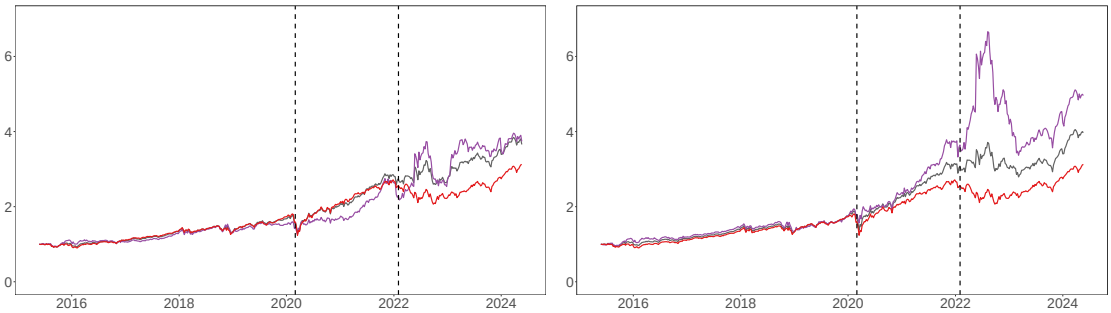
(a) Benchmark portfolio wealth curves.

(b) Portfolio wealth curves with $N = 12$ weeks.



(c) Portfolio wealth curves with $N = 24$ weeks.

(d) Portfolio wealth curves with $N = 52$ weeks.



The data is weekly from June 1, 2015, to May 21, 2024, with three periods: normal (June 1, 2015 – March 12, 2020), COVID-19 (March 12, 2020 – February 24, 2022), and Russia-Ukraine (February 24, 2022 – May 21, 2024).

3.5 Conclusion

The objective of this chapter is to revisit whether energy futures can add value to an equity portfolio in light of recent data covering the COVID-19 pandemic and the Russia-Ukraine war, two crises that have profoundly influenced the energy landscape. While previous works covering these crises have focused primarily on the increased volatility and changing correlations during these crisis periods, our paper builds upon existing literature by assessing the financial performance of equity portfolios enhanced with energy futures. Moreover, while several studies omit electricity as a viable investment opportunity, we also include electricity along with crude oil and natural gas. This paper proposes an empirically-based mean-variance asset allocation strategy to enhance the risk-adjusted return profile of a stock portfolio using energy futures. Our model has the ability to capture the non-stationary dynamics inherent in the underlying data and adjust portfolio positions accordingly. Our results demonstrate improvements in the out-of-sample Sharpe and Sortino ratios relative to the static S&P 500 benchmark portfolio in each period considered, particularly during crisis periods. Furthermore, imposing moderate bounds on energy futures positions stabilizes allocations and substantially reduces risk. Hence, despite the increased financialization of commodity markets in recent years, energy futures can improve the out-of-sample risk-return performance of a conventional equity portfolio. A contributing factor is the divergent behaviours of the various energy commodities across the different periods and crises. For example, electricity and natural gas, which are highly correlated in regular time periods, have opposite trends during the Russia-Ukraine conflict.

4 On the use of quantile-regression-based DEXA phenotypes to assess health risks

4.1 Introduction

According to the World Health Organization, as of 2022, 1 in 8 people globally are living with obesity. In 2019, the obesity epidemic and its associated health risks caused an estimated 5 million deaths from non-communicable diseases. The global costs of this epidemic are expected to hit US$ 3 trillion per year by 2030. The importance of accurately identifying patients at risk of obesity-related health effects is thus paramount.

Using body mass index (BMI) as a proxy, obesity is generally defined as a BMI $\geq 30\text{kg}/\text{m}^2$. However, BMI does not differentiate between muscle mass and fat mass, and its limitations are well described in Rothman (2008); Frankenfield et al. (2001); Nuttall (2015). Dual energy X-ray absorptiometry (DEXA) is thus often used for assessing body composition with higher precision as it can differentiate between fat mass and muscle mass.

Following this approach, Prado et al. (2014) constructed fat and muscle mass phenotypes using DEXA data from a representative sample of the US adult population. Simultaneously assessing fat mass and muscle mass can help to identify nuanced health risks, thus allowing for better risk stratification. For example, two individuals with the same BMI might have different fat-to-muscle ratios, resulting in different health outcomes. Prado et al. thus developed four mutually exclusive DEXA-derived phenotypes based on whether an individual was above or below the median of DEXA-measured adiposity and muscle mass indices for their respective sex and age reference curves. The lambda-mu-sigma methodology (LMS; Appendix C) used by Prado et al. has been well-validated in the literature (Cole and Green (1992); Flegal and Cole (2013)) and was originally developed to construct growth charts for children. However, previous work by Kakinami et al. (2022) has shown that median-split phenotypes are not more effective than BMI at identifying cardiometabolic risk.

These studies have some limitations. Using the medians of the fat-mass and muscle-mass indices to construct the DEXA-derived phenotypes might be too simplistic. Additionally, the LMS methodology does not capture the kurtosis, which is an important measure to help identify patients lying in the tails of the distributions. Thus, a more rigorous exploration of the DEXA-derived phenotype clustering space is needed. For instance, quantile regression may better capture the full spectrum of the data distribution, including the tail behavior. Lastly, these studies only focus on metabolic risk factors, such as cardiovascular disease and diabetes. However, in Dixon (2010), obesity has been shown to have an effect upon various health outcomes, including physical and mental health, quality of life, and comorbidities. The consideration of other health risks must therefore be investigated, as the effects of obesity on health are widespread. The objective of this study was thus to develop a DEXA-derived phenotype classification system that:

- captures kurtosis,
- and considers additional centile cut-offs beyond the median.

We then compare the performance of our models to standard adiposity measures, such as BMI and waist circumference. This study focused on metabolic syndrome as the primary health outcome,

but additionally assessed other health risks, such as depression, sleep disorders, general health, instrumental activities of daily living, and comorbidities.

The remainder of this chapter is organized as follows. In Section 4.2, we describe the data. In Section 4.3, we present the different exposures. In Section 4.4, we introduce the health outcomes. In Section 4.5, we outline the covariates. In Section 4.6, we provide model specifications. In Section 4.7, we detail the diagnostic accuracy metrics used in this study. Section 4.8 is a statistical analysis. In Sections 4.9 and 4.10, we report on our findings together with a related discussion. In Section 4.11, we conclude.

4.2 Description of the data

Data was from the 2011-2018 waves of the National Health and Nutrition Examination Survey (NHANES). NHANES was designed to evaluate the health and nutritional status of both children and adults using a cross-sectional representative sample of the US general population, combining both interviews and physical examinations (NHANES (2024); Curtin et al. (2012)). The first stage in the complex survey design consisted of selecting primary sampling units (PSUs), which were selected with a probability in proportion to a measure of size (PPS). The measure of size (MOS) is a weighted average of population counts. The weights give a higher probability of selection to PSUs containing larger proportions of demographic subgroups that were selected for oversampling (where oversampling was used to increase the reliability and precision of estimates of health status indicators for population subgroups identified by NHANES). In the second stage, the sampled PSUs were divided into segments, which were selected with PPS. In the third stage, dwelling units within each segment were listed and then randomly sampled. In the fourth and final stage, individuals from a list of all people residing in the selected dwelling units were randomly sampled. Each annual sample is nationally representative. However, due to time and cost constraints, NHANES can only survey a relatively small number of participants per year. Single-year data might thus lead to unstable parameter estimates and may further increase the possibility of patient disclosure, hence data is publicly released according to a two-year cycle. As previously discussed, the participants in NHANES have an unequal probability of selection, with an additional adjustment for sample person non-response. Sample weights were thus used to produce unbiased national estimates. Participants provided written consent; the study was approved by the National Center for Health Statistics' institutional review board.

The data was filtered according to the pipeline depicted in the flowchart in Figure 21. As the Prado et al. methodology identified phenotypes exclusively among adults, participants below the age of 20 were excluded. The 2011-2018 waves of NHANES include 39156 participants, of which 4426 were retained for our study.

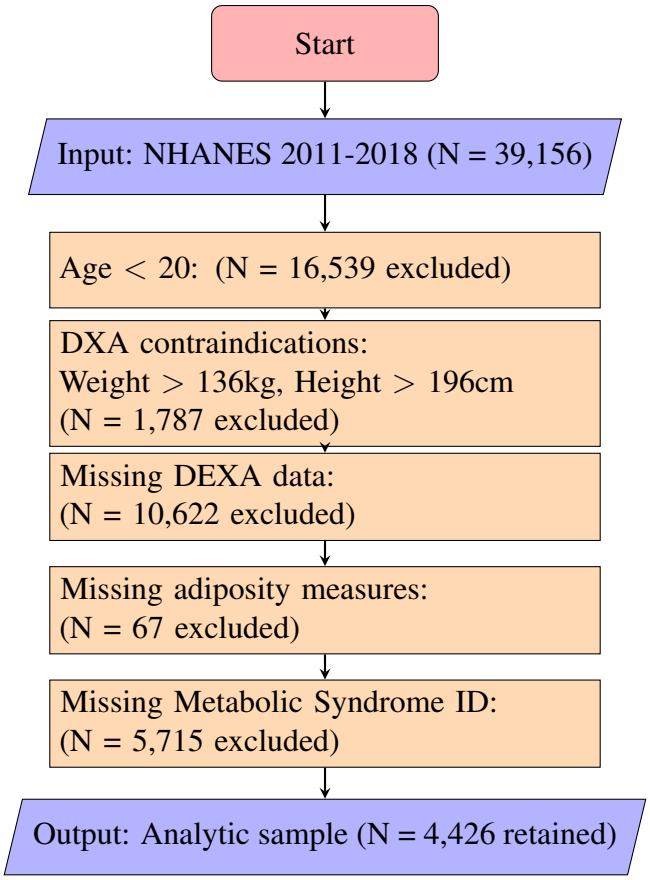


Figure 21: Flowchart of exclusions for the analytic study sample, NHANES (2011-2018)

4.3 Measures

4.3.1 Exposures

To obtain lean and fat mass, a whole body DEXA scan (Hologic QDR 4500A) was used. To ensure the quality of the DEXA scans administered at the NHANES mobile examination centers, the Hologic Anthropomorphic Spine Phantom was scanned daily for accurate calibration of the densitometer (CDC and NCHS (2021)). The fat mass index (FMI) and appendicular skeletal mass index (ASMI) were computed by dividing the DEXA-measured fat and appendicular skeletal mass (i.e. lean soft tissue from the arms and legs) in kilograms, respectively, by the square of the height in meters (Heymsfield et al. (1990); Hattori et al. (1997)).

Trained examiners measured waist circumference (in cm) in accordance with standard procedures. Height (cm) and weight (kg) were measured using a stadiometer and an electronic digital scale, respectively. BMI (kg/m^2) and waist-to-height ratio were calculated with these measured values (Ashwell et al. (2012)). Total percent fat (%) was determined using DEXA data.

4.4 Health outcomes

4.4.1 Metabolic syndrome

Cardiometabolic risk factors include high-density lipoprotein (HDL, mg/dl), triglycerides (TG, mg/dl), fasting plasma glucose (FPG, mg/dl), and blood pressure expressed as systolic and diastolic pressures (mmHg).

As described in the NHANES laboratory procedures manual (NHANES (2016)), a phlebotomist used venipuncture to obtain blood samples from participants in the mobile examination centers, which were then transported to Johns Hopkins University Lipoprotein Analytical Laboratory for analysis. Serum samples were stored at -20 °C until being shipped, or at -80 °C if they were stored for over a month. Fasting plasma glucose was measured using a hexokinase enzymatic method in a central laboratory. Systolic and diastolic blood pressure (SBP and DBP, respectively) were measured on the participant's right arm using a sphygmomanometer after spending 5 minutes in a resting, seated position. Since multiple SBP and DBP readings were recorded, details on how the averages were calculated are available in the documentation on the NHANES website.

The criteria for metabolic syndrome (MetS), according to the National Cholesterol Education Program (NCEP) guidelines, requires that participants present with at least three of the following factors: waist circumference $>$ 88cm for females or $>$ 102cm for males, blood pressure $\geq$ 135mmHg (systolic) or $\geq$ 85mmHg (diastolic), TG $\geq$ 150mg/dl, HDL $<$ 50mg/dl for females or $<$ 40mg/dl for males, or FPG $\geq$ 100mg/dl (Miller et al. (2014)).

4.4.2 Depression

Mental-health-related questionnaire data was obtained from the mental health depression screener (DPQ). The following nine symptoms from the DPQ questionnaire were used for this health outcome: (1) have little interest in doing things, (2) feeling down, depressed, or hopeless, (3) trouble sleeping or sleeping too much, (4) feeling tired or having little energy, (5) poor appetite or overeating, (6) feeling bad about yourself, (7) trouble concentrating on things, (8) moving or speaking slowly or too fast, and (9) thought you would be better off dead. For each symptom, points range from 0 to 3, corresponding to *not at all*, *several days*, *more than half the days*, and *nearly every day*. A score $\geq$ 10 was used to identify those with major depression in accordance with guidelines that demonstrate this has a sensitivity and specificity of 88% (Kroenke et al. (2001)).

4.4.3 Short sleep

The *sleep hours* variable from the sleep disorders questionnaire was used for this health outcome. Participants self-reported their average amount of sleep during workdays, in hours. Short sleep was defined as less than 6 hours per night (Beccuti and Pannain (2011)).

4.4.4 General health

The *general health condition* of a participant was obtained through the current health status questionnaire, and was further dichotomized for modeling purposes. We created two categories to represent high and low general health conditions: *excellent*, *very good*, and *good* factor levels correspond to high, whereas *fair* and *poor* correspond to low (Reichmann et al. (2011)).

4.4.5 Physical functioning

We used the activities of daily living (ADL) and the instrumental activities of daily living (IADL) from the physical functioning questionnaire to assess a participant's physical functioning (Jindai et al. (2016)). The ADL domain includes the following functional limitations: (1) getting in and out of bed, (2) using a fork or knife; or drinking from a cup, (3) walking between rooms on the same floor, and (4) dressing yourself. The IADL domain includes: (1) house chores, (2) managing money, and (3) preparing meals. Responses were coded as 0 if the participant reported performing an activity with *no difficulty*; responses were coded as 1 if the participant reported *some difficulty*, *much difficulty*, or *unable to do*. For each of the ADL and IADL domains, a response equal to 1 in any of the associated functional limitations was interpreted as corresponding to a physical functioning deficiency (Jindai et al. (2016)).

4.4.6 Comorbidities

We examined the presence of comorbidities (i.e., presenting with two or more illnesses) amongst the sample participants (King et al. (2018)). In the medical conditions questionnaire, we included the following yes/no items: doctor ever said you had (1) asthma, (2) arthritis, (3) gout, (4) thyroid problems, (5) liver conditions, (6) cancer or malignancy, (7) cardiovascular disease or CVD (which includes congestive heart failure, coronary heart disease, angina pectoris, heart attack or stroke), and (8) chronic obstructive pulmonary disease or COPD (which includes emphysema or chronic bronchitis). The presence of at least two medical conditions was necessary to receive a positive comorbidity classification.

4.5 Covariates

4.5.1 Demographic characteristics

Sex, ethnicity, marital status, and education level were self-reported during the interview section of the NHANES survey. The ratio of family income to poverty was calculated based on self-reported previous year's household income according to the Department of Health and Human Services guidelines, with adjustments for household size, geographic location, and inflation.

4.6 Model specification

4.6.1 Median-split phenotypes

Using the LMS methodology described in Cole and Green (1992); Flegal and Cole (2013), Prado et al. developed four mutually exclusive phenotypes based on whether an individual was above or below the median of DEXA-measured adiposity and muscle mass indices for their respective sex and age reference curves: high-adiposity, high-muscle (HA_{median}-HM_{median}, $\geq$ 50th centile FMI and $\geq$ 50th centile ASMI), high-adiposity, low-muscle (HA_{median}-LM_{median}, $\geq$50th centile FMI and $<$ 50th centile ASMI), low-adiposity, high-muscle (LA_{median}-HM_{median}, $<$ 50th centile FMI and $\geq$ 50th centile ASMI), and finally low-adiposity, low-muscle (LA_{median}-LM_{median}, $<$ 50th centile FMI and $<$ 50th centile ASMI).

4.6.2 Quantile regression phenotypes

As an extension to ordinary least squares regression, quantile regression was used to robustly estimate the conditional quantile of a response variable, across different values of the predictor variables. It has been used, for example, to develop growth charts to help identify abnormal growth. The τth conditional quantile of a random variable Y given a K-dimensional random vector $\mathbf{X}$ is defined as

$$Q_{Y|\mathbf{X}} = \inf\{y : F_{Y|\mathbf{X}}(y) \geq \tau\},$$

where $F_{Y|\mathbf{X}}(\cdot)$ denotes a conditional cumulative distribution function (CDF). In the case of quantile regression, the τth conditional quantile is expressed as a linear function of the form

$$Q_{Y|\mathbf{X}} = \mathbf{X}\beta_\tau,$$

for some K-dimensional parameter vector β_τ. The optimal set of parameters can be obtained by solving

$$\beta_\tau = \underset{\beta \in \mathbb{R}^K}{\arg\min}\, \mathbb{E}[\rho_\tau(Y, \mathbf{X}\beta)],$$

where ρ_τ is the pinball loss function, defined as

$$\rho_\tau(y, \hat{y}) = \begin{cases} (y - \hat{y})\tau, & \text{if } y \geq \hat{y} \\ (\hat{y} - y)(1 - \tau), & \text{otherwise} \end{cases}$$

where y is the realization of the random variable Y and $\hat{y}$ is the quantile forecast. We can then replace the theoretical expectation with the empirical mean computed using the observed data, $\{(y_i, \mathbf{x}_i)\}_{i=1}^n$, to estimate

$$\hat{\boldsymbol{\beta}}_\tau = \underset{\beta \in \mathbb{R}^K}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \rho_\tau\big(y_i, \mathbf{x}_i^\top \beta\big).$$

As a first step, hyperparameter tuning was used to select an adequate set of sex-, ethnicity-, and age-stratified lower and upper centile cut-offs (that is, conditional quantiles) in order to construct the DEXA-derived phenotypes. The 25% and 75% centiles were selected as the cut-offs for both the FMI and ASMI indices to ensure a reasonable number of observations in each phenotype. The resulting DEXA-derived phenotypes will be referred to hereafter as quantile regression (QR) phenotypes.

4.7 Diagnostic accuracy metrics

Several metrics can be used to address the merits of a diagnostic test (Šimundić (2009)). The sensitivity of a diagnostic test is the proportion of true positive subjects with the disease amongst the total group of subjects with the disease. This can be interpreted as the probability of obtaining a positive test result (T+), given that the subject has the disease (D+), or (T+ |D+). On the other hand, the specificity of a diagnostic test is the proportion of true negative subjects without the disease amongst the total group of subjects without the disease. This can be understood as the probability of getting a negative test result (T-), given that the subject does not have the disease (D-), or (T- |D-). Sensitivity and specificity do not depend upon disease prevalence.

From a clinical standpoint, however, the point of view taken by physicians for the purpose of diagnosis is the presence/absence of disease given a positive/negative test result, that is: (D+ |T+) and (D- |T-). The predictive values are thus of greater interest (Baeyens et al. (2019)). The positive predictive value (PPV) is the probability that a subject has the disease, given a positive test result, or (D+ |T+). The negative predictive value (NPV) is the probability that a subject does not have the disease, given a negative test result, or (D- |T-). Unlike sensitivity and specificity, however, predictive values depend upon disease prevalence. PPV increases with higher prevalence, whereas NPV is decreasing.

A metric that is both clinically relevant and independent of disease prevalence is thus ideal. We therefore additionally considered the likelihood ratio. The positive likelihood ratio (LR+) refers to how many times more likely it is to obtain a positive test result amongst subjects with the disease, relative to those without, or (T+ |D+)/(T+ |D-). LR+ is a good indicator for ruling-in a disease, with larger values providing greater diagnostic ability (i.e. strong diagnostic tests have LR+ > 10). LR+ can also be expressed as sensitivity/(1-specificity). Conversely, the negative likelihood ratio (LR-) tells us how likely it is to get a negative test result amongst subjects with the disease, relative to those without, or (T- |D+)/(T- |D-). LR- is a good indicator for ruling-out a disease, with lower values giving us increased diagnostic ability (i.e. strong diagnostic tests have LR- < 0.1). LR- can also be defined as (1-sensitivity)/specificity. In general, likelihood ratios equal to 1 are not diagnostic.

4.8 Statistical analysis

Statistical analyses were performed using `R` and packages `survey`, `srvyr`, and `quantreg`. The complex survey design accounts for sampling weights, primary sampling units, as well as strata. The quantile regression and logistic regression models were trained using the in-sample training set, consisting of the NHANES waves 2011-2016 (n = 3,524). Performance was then assessed using the NHANES waves 2017-2018 (n = 902) as an out-of-sample validation set. In addition to this, the quantile regression model further incorporated bootstrapped replicates. The predicted quantiles were used to segment the FMI/ASMI coordinate space into nine distinct phenotypes, while controlling for age, gender, and race. The FMI/ASMI distributions are graphically presented using histograms, whereas the resulting QR phenotype clusters are displayed with scatterplots. We additionally looked at the overlap between the AvgA_{QR}-AvgM_{QR} phenotype with the median-split phenotypes in a profile analysis, and compared the adiposity profiles amongst the patients in the overlapping regions. This analysis was performed to assess whether the median-split model incorrectly classified patients with similar adiposity profiles into different phenotypes. After constructing the QR phenotypes, these clusters were then used to predict the presence of the health outcome of interest by means of logistic regression. The logistic regression model can be defined as:

$$\log \frac{\mathbb{P}(Y=1|\mathbf{X})}{\mathbb{P}(Y=0|\mathbf{X})} = \mathbf{X}\beta, \tag{11}$$

such that the β coefficients correspond to the log-odds. The odds ratios are then given by $\exp(\beta)$. An odds ratio can be interpreted as the $\frac{\text{odds of disease for exposed}}{\text{odds of disease for unexposed}}$, such that the odds of an event is given by $\frac{\mathbb{P}(\text{event})}{1-\mathbb{P}(\text{event})}$. In the case of a single categorical predictor, the reference group is unexposed. Performance was then assessed using LR+ and LR-, and AIC. Additional health

outcomes, including depression, short sleep, general health, physical functioning, and comorbidities were also investigated. Finally, the predictive power of other adiposity measures was explored, including BMI, waist circumference, and total fat %.

4.9 Results

4.9.1 Descriptive statistics

Descriptive statistics did not vary significantly across the in-sample (NHANES 2011-2016) and out-of-sample (NHANES 2017-2018) datasets (data not shown). In Table 1, the weighted descriptive statistics for NHANES 2011-2016 and 2017-2018 are provided. In either dataset, the LA_{QR}-LM_{QR} group is characterized by the lowest BMI, lean mass index (i.e., ASMI / squared height, in kg/m^2, as outlined in Minetto et al. (2021)), and waist circumference. In contrast, the HA_{QR}-HM_{QR} group exhibits the highest BMI, lean mass index, and waist circumference; and is predominantly female. Additionally, the LA_{QR}-HM_{QR} group stands out for its relatively higher levels of education, suggesting potential socioeconomic differences across the groups.

Recall that the lower and upper centile cut-offs for the QR phenotypes were respectively set at 25% and 75% (Table 14). For the purpose of illustration, the LA_{QR}-LM_{QR} phenotype has FMI and ASMI values that both fall below the 25% centile cut-offs; the $AvgA_{QR}$-$AvgM_{QR}$ phenotype has FMI and ASMI values that both fall between the 25% and 75% centile cut-offs; and the HA_{QR}-HM_{QR} phenotype has FMI and ASMI values that both fall above the 75% centile cut-offs.

Figure 22 illustrates the overlap between the QR and median-split phenotypes in a representative sub-sample from NHANES 2011-2016 and 2017-2018. The scatter plots show a positive association between ASMI and FMI, where increased fat mass is often accompanied by increased muscle mass (and vice versa). The profile analysis (Table 15) provides further context; the majority of the patients classified as $AvgA_{QR}$-$AvgM_{QR}$ by the QR model are in the two extremes of the median-split phenotypes (i.e. LA_{median}-LM_{median} and HA_{median}-HM_{median}).

4.9.2 Model fit and performance

Based on in-sample AIC, the BMI model was optimal for MetS, general health, and IADL, while the WC model was best for depression, short sleep, and comorbidity (Table 16). Examining out-of-sample performance, the BMI and WC models tend to have the optimal LR+ and LR- values. The QR model generally shows a better (lower) AIC than the median-split model across outcomes, except for short sleep and IADL. Out-of-sample, the QR model has a better (higher) LR+ than the median-split model for MetS and comorbidity (with short sleep being ambiguous, as both models show LR+ values ≤ 1). However, the QR model consistently underperforms in LR- compared to the median-split model (with short sleep again ambiguous as LR- ≥ 1 for both models).

4.10 Discussion

Comparing the QR model to the median-split model, we find that the QR model generally has a better in-sample fit than the median-split model for most outcomes, except short sleep and IADL. As the data is densely populated close to the center, the median-split model thereby assigns nearby points to different classes. Indeed, for points close to the center, the differences in adiposity measures

across the median-split phenotypes are minimal, both in- and out-of-sample, thus underscoring a limitation of the median-split model. Previous studies by Rousseeuw and Hubert (2011); Huber and Ronchetti (2011) have discussed the trade-off between the robustness and efficiency of the median. The efficiency trade-off implies that when the data are densely concentrated, the median may fail to detect differences between nearby points, which can potentially result in less reliable boundaries for the phenotypes. In contrast, the QR model mitigates this inconsistency. For instance, it defines its AvgA_{QR}-AvgM_{QR} phenotype to encompass data points that, under Prado's classification, would fall into various phenotype categories. Nevertheless, as the data become sparser further from the center, the discrepancies between the QR and median-split models diminish. Future work may therefore address whether having the nine phenotypes identified using the QR model is necessary, or whether there exists an optimal number of phenotypes falling somewhere between four (as suggested in the median-split model) and nine.

In addition, the QR model has relatively larger out-of-sample LR+ values for METS and co-morbidity, thus indicating a relative strength in confirming these health outcomes. However, the median-split model achieves better out-of-sample LR- values, hence indicating an advantage in ruling out these outcomes compared to the QR model. The trade-off between the LR+ and LR- values are a direct consequence of the way in which the thresholds are set when constructing the phenotypes. That is, the higher specificity (fewer false positives) of the QR model comes at the expense of reduced sensitivity (more false negatives), thereby improving the LR+ but deteriorating the LR-. To help make this point clear, consider the generic HA-HM phenotype as an example. The QR model uses the 75% centiles as cut-offs, whereas the median-split model uses the medians. The higher threshold imposed by the QR model results in fewer patients being classified as HA_{QR}-HM_{QR} relative to $\text{HA}_{\text{median}}$-$\text{HM}_{\text{median}}$, which ultimately results in fewer false positives but more false negatives when predicting the associated health outcomes. The relative importance of LR+ versus LR- is contingent upon the clinical context, as well as on the primary goal of the diagnostic test (Jaeschke et al. (1994)). For example, if the goal is to ensure that the patients identified as high-risk are indeed more likely to have the disease, as in the case of high-risk treatments, then a higher LR+ is preferable. Conversely, if there is substantial danger in leaving the target disease undiagnosed and therefore untreated, then a lower LR- is advantageous, even if it means treating the disease unnecessarily in some patients. In this context, whether higher LR+ or lower LR- is more important is nevertheless unclear and therefore requires further investigation.

Our study suggests that the HA_{QR}-HM_{QR} phenotype is associated with higher metabolic and general health risks. In addition, we find that the LA_{QR}-HM_{QR} phenotype may act as a protective factor. This highlights the importance of using body composition profiles to enhance the evaluation of health risks. For example, in Zamboni et al. (2008), high fat mass accompanied by low muscle mass has been shown to contribute to disability, morbidity, and mortality, especially in older adults.

The BMI and WC models consistently demonstrate the best in-sample fit and out-of-sample performance among the models evaluated. However, while BMI and WC may perform well overall, they are likely to misclassify certain groups, as discussed in Nuttall (2015); Ode et al. (2007); Zamboni et al. (2008); Després (2012). In the elderly, for example, sarcopenia can result in ambiguous BMI values, while athletes with high lean mass can have an elevated BMI despite their low body fat levels. In addition, BMI cutoffs may not be widely applicable across different racial and ethnic sub-populations. Hence, although QR may not outperform BMI and WC in general, it could potentially be better in specific demographics where these traditional measures are prone to fail. Future work addressing this issue is therefore needed.

As an alternative to DEXA, other methods to assess body composition or adiposity should be explored, such as magnetic resonance imaging (MRI), computed tomography (CT), three-dimensional optical imaging (3DO), ultrasound, and bioelectrical impedance analysis (BIA), to list a few. However, as explained in Lukaski (1987); Tewari et al. (2018); Thomas et al. (2025), there are many considerations that must be taken into account when selecting a method for measuring body composition, such as cost, portability, safety, and accuracy. For example, BIA is more cost-effective relative to DEXA, is safe for all populations, but has nevertheless been shown to underestimate fat-free mass. In the case of CT scans, their relatively higher levels of radiation exposure limits their use to patients with illness. An interesting alternative is the smartphone camera, which is relatively inexpensive, safe, portable, and agrees well with DEXA when used in conjunction with machine learning models.

This study is not without limitations. The HA_{QR}-LM_{QR} and LA_{QR}-HM_{QR} phenotypes had limited representation, containing a relatively small number of patients. Hence, despite NHANES being a large representative sample of the US adult population, focusing on a larger sample with more people at the extremes would help to better validate performance in the tails of the distribution. Moreover, as an alternative to using quantile regression, clustering techniques could also be used to generate DEXA-derived phenotypes. The machine learning literature includes several different approaches, such as K-means clustering. However, compared to quantile regression, clustering models are often more sensitive to outliers, which can distort both cluster assignments and the resulting group structure (Nowak-Brzezińska and Gaibei (2022); Abdussamad and Inayat (2024)). The cross-sectional design used in our study further limits our ability to establish causality. Longitudinal studies would therefore be required for causal inference and to determine whether changes in phenotypes can predict health outcomes (Prado et al. (2014)). Other adiposity metrics might also be explored. For example, a prior study by Woolcott and Bergman (2018) used a sex-stratified linear combination of the height-to-waist circumference to predict whole-body fat percentage, resulting in improved accuracy relative to BMI. Lastly, this study focused on adults at least 20 years of age. Moreover, the maximum age observed in our sample is 80 years. The results of our study cannot therefore be extended to the general US population, particularly to younger and older age groups. Thus, further research is needed to better understand the risks faced by vulnerable subpopulations, such as older adults, where sarcopenia and sarcopenic obesity are known to have harmful health effects (Wannamethee and Atkins (2015)).

4.11 Conclusion

In this paper, we used quantile regression as an alternative to the LMS methodology to construct DEXA-derived phenotypes, as it was better suited to capture high-risk or atypical patients. The QR model was more effective than the median-split model in correctly identifying patients with MetS and comorbidities; but it underperformed in identifying individuals without these conditions. The BMI and WC models consistently demonstrated the best overall performance among the models evaluated. However, even though QR may not outperform BMI and WC in general, whether it could be better among vulnerable subpopulations should be explored. Whether the classification performances diverge in longitudinal studies should also be investigated.

Table 13: Weighted descriptive statistics.

NHANES 2011-2016[a]	$AvgA_{QR}$-$AvgM_{QR}$ (N = 1,103)	$AvgA_{QR}$-HM_{QR} (N = 255)	$AvgA_{QR}$-LM_{QR} (N = 410)	HA_{QR}-$AvgM_{QR}$ (N = 263)	HA_{QR}-HM_{QR} (N = 559)	HA_{QR}-LM_{QR} (N = 10)	LA_{QR}-$AvgM_{QR}$ (N = 327)	LA_{QR}-HM_{QR} (N = 7)	LA_{QR}-LM_{QR} (N = 590)
Gender									
Female	53%	36%	53%	37%	57%	8%	52%	28%	49%
Male	47%	64%	47%	63%	43%	92%	48%	72%	51%
Age, years	39.2 (11.5)	40.2 (10.1)	39.2 (12.2)	40.3 (12.3)	39.0 (11.2)	40.0 (7.7)	39.0 (11.8)	42.7 (12.9)	39.6 (12.8)
Race/ethnicity									
Asian	7%	0%	15%	6%	1%	30%	4%	0%	10%
Black	10%	17%	4%	9%	18%	0%	14%	34%	9%
Hispanic	20%	11%	26%	16%	13%	0%	12%	21%	28%
White	64%	72%	56%	69%	67%	70%	70%	46%	53%
Marital status									
Divorced/widowed/separated	13%	8%	10%	16%	12%	8%	9%	15%	15%
Married/in a relationship	66%	68%	63%	57%	63%	40%	65%	76%	60%
Single (never married)	21%	24%	27%	27%	25%	51%	26%	9%	25%
PIR < 1.30, %	22%	19%	28%	23%	26%	21%	21%	16%	34%
At least high-school education, %	86%	86%	80%	90%	88%	77%	89%	93%	78%
Adiposity measures									
BMI	27.7 (3.1)	30.4 (2.8)	25.1 (2.8)	32.6 (3.4)	37.5 (5.0)	29.5 (2.7)	23.3 (2.6)	27.3 (3.1)	21.5 (2.6)
Lean mass index, kg/m^2	17.7 (2.2)	20.6 (2.1)	15.6 (1.9)	19.0 (2.0)	21.7 (2.4)	17.7 (1.8)	16.8 (2.4)	21.2 (3.3)	15.1 (2.1)
Total % fat	33.3 (6.9)	29.6 (6.6)	35.0 (6.9)	39.0 (6.6)	39.6 (6.8)	37.6 (3.8)	25.2 (5.9)	19.5 (5.5)	26.6 (7.2)
Trunk % fat	32.3 (6.6)	29.5 (6.3)	33.0 (6.5)	38.8 (5.9)	39.3 (6.0)	38.0 (2.6)	23.0 (5.8)	18.0 (5.3)	24.1 (7.1)
Waist circumference, cm	95.6 (8.3)	102.3 (7.8)	88.7 (7.7)	109.1 (7.9)	117.7 (10.8)	103.8 (6.1)	84.1 (7.8)	94.9 (12.3)	79.7 (7.9)
Waist-to-height	0.6 (0.1)	0.6 (0.1)	0.5 (0.1)	0.6 (0.1)	0.7 (0.1)	0.6 (0.0)	0.5 (0.0)	0.5 (0.1)	0.5 (0.1)

NHANES 2017-2018[a]	$AvgA_{QR}$-$AvgM_{QR}$ (N = 283)	$AvgA_{QR}$-HM_{QR} (N = 58)	$AvgA_{QR}$-LM_{QR} (N = 113)	HA_{QR}-$AvgM_{QR}$ (N = 80)	HA_{QR}-HM_{QR} (N = 142)	HA_{QR}-LM_{QR} (N = 3)	LA_{QR}-$AvgM_{QR}$ (N = 80)	LA_{QR}-HM_{QR} (N = 11)	LA_{QR}-LM_{QR} (N = 132)
Gender									
Female	48%	61%	42%	34%	59%	32%	49%	42%	57%
Male	52%	39%	58%	66%	41%	68%	51%	58%	43%
Age, years	37.9 (11.1)	42.7 (11.3)	38.1 (12.3)	39.1 (12.0)	37.2 (10.5)	41.1 (4.2)	39.5 (12.4)	40.6 (11.4)	41.9 (13.4)
Race/ethnicity									
Asian	9%	1%	16%	11%	2%	58%	5%	0%	4%
Black	13%	11%	6%	7%	17%	0%	12%	28%	6%
Hispanic	22%	16%	27%	19%	17%	0%	13%	13%	28%
White	55%	71%	51%	62%	64%	42%	71%	59%	62%
Marital status									
Divorced/widowed/separated	12%	21%	10%	22%	10%	0%	12%	31%	15%
Married/in a relationship	65%	63%	50%	53%	61%	58%	64%	25%	59%
Single (never married)	22%	16%	40%	25%	28%	42%	24%	45%	26%
PIR < 1.30, %	22%	15%	17%	28%	22%	42%	20%	4%	20%
At least high-school education, %	90%	92%	88%	91%	91%	100%	89%	100%	89%
Adiposity measures									
BMI	27.6 (2.9)	32.1 (3.2)	25.4 (2.6)	33.2 (3.6)	38.1 (4.5)	27.8 (0.7)	22.9 (2.5)	27.6 (4.0)	21.5 (2.6)
Lean mass index, kg/m^2	17.8 (2.4)	20.6 (2.3)	16.0 (2.0)	19.7 (2.2)	21.7 (2.4)	15.6 (1.6)	16.8 (2.1)	20.5 (3.9)	14.7 (2.0)
Total % fat	32.5 (7.2)	33.1 (7.4)	34.1 (6.9)	38.2 (7.1)	40.4 (6.5)	41.4 (6.9)	23.4 (5.2)	23.2 (5.3)	27.9 (7.0)
Trunk % fat	31.8 (6.7)	31.4 (6.5)	33.2 (6.5)	37.8 (5.7)	40.1 (6.0)	42.5 (5.7)	21.2 (5.0)	22.5 (4.8)	25.0 (6.8)
Waist circumference, cm	94.9 (7.8)	104.1 (8.8)	89.7 (7.1)	108.9 (7.8)	119.1 (10.6)	101.8 (1.7)	82.1 (7.6)	90.7 (11.7)	79.0 (7.5)
Waist-to-height	0.6 (0.1)	0.6 (0.1)	0.5 (0.1)	0.6 (0.1)	0.7 (0.1)	0.6 (0.0)	0.5 (0.0)	0.5 (0.1)	0.5 (0.0)

The HA_{QR}-LM_{QR} and LA_{QR}-HM_{QR} phenotypes contain too few observations to confidently interpret the findings.
PIR = Poverty Income Ratio, BMI = Body Mass Index.
[a]Mean (standard error).

Table 14: QR phenotypes, based on FMI and ASMI.

FMI LCF	FMI UCF	ASMI LCF	ASMI UCF	QR phenotype
$-$	$-$	$-$	$-$	LA-LM
$+$	$-$	$-$	$-$	AvgA-LM
$+$	$+$	$-$	$-$	HA-LM
$-$	$-$	$+$	$-$	LA-AvgM
$+$	$-$	$+$	$-$	AvgA-AvgM
$+$	$+$	$+$	$-$	HA-AvgM
$-$	$-$	$+$	$+$	LA-HM
$+$	$-$	$+$	$+$	AvgA-HM
$+$	$+$	$+$	$+$	HA-HM

Notes: Data was from NHANES (2011-2018). FMI LCF is the lower centile cut-off for the FMI index, FMI UCF is the upper centile cut-off for the FMI index, ASMI LCF is the lower centile cut-off for the ASMI index, and ASMI UCF is the upper centile cut-off for the ASMI index. LA is low-adiposity, AvgA is average adiposity, HA is high adiposity, LM is low muscle mass, AvgM is average muscle mass, and HM is high muscle mass. For ease of visual representation, $-$ is used to denote $<$ and $+$ is used to denote $>$.

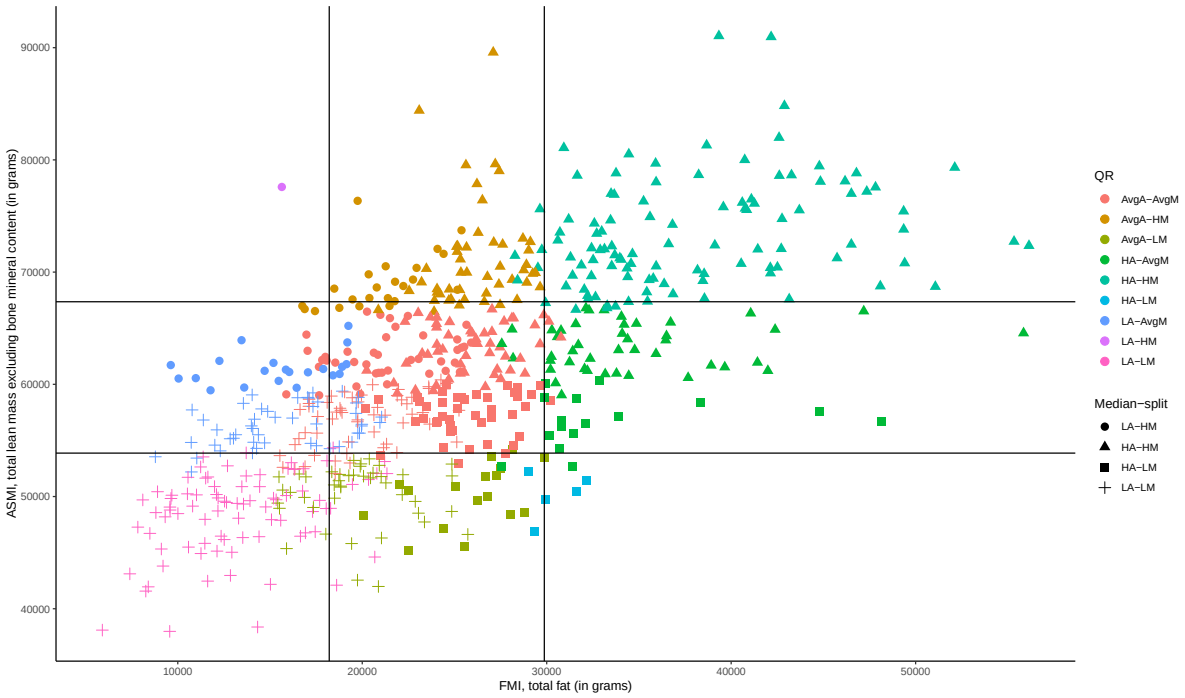
Table 15: Profile analysis of the AvgA_{QR}-AvgM_{QR} phenotype.

	NHANES 2011-2016 Median-split phenotypes				NHANES 2017-2018 Median-split phenotypes			
Adiposity measures	LA-HM (N = 87)	HA-HM (N = 249)	HA-LM (N = 82)	LA-LM (N = 230)	LA-HM (N = 16)	HA-HM (N = 51)	HA-LM (N = 14)	LA-LM (N = 50)
BMI	27	29	28	26	28	29	28	26
Waist circumference, cm	96	101	100	92	95	104	99	94
Waist-to-height	0.54	0.57	0.57	0.53	0.54	0.59	0.57	0.54
Lean mass index, kg/m^2	20	20	19	18	21	20	19	19

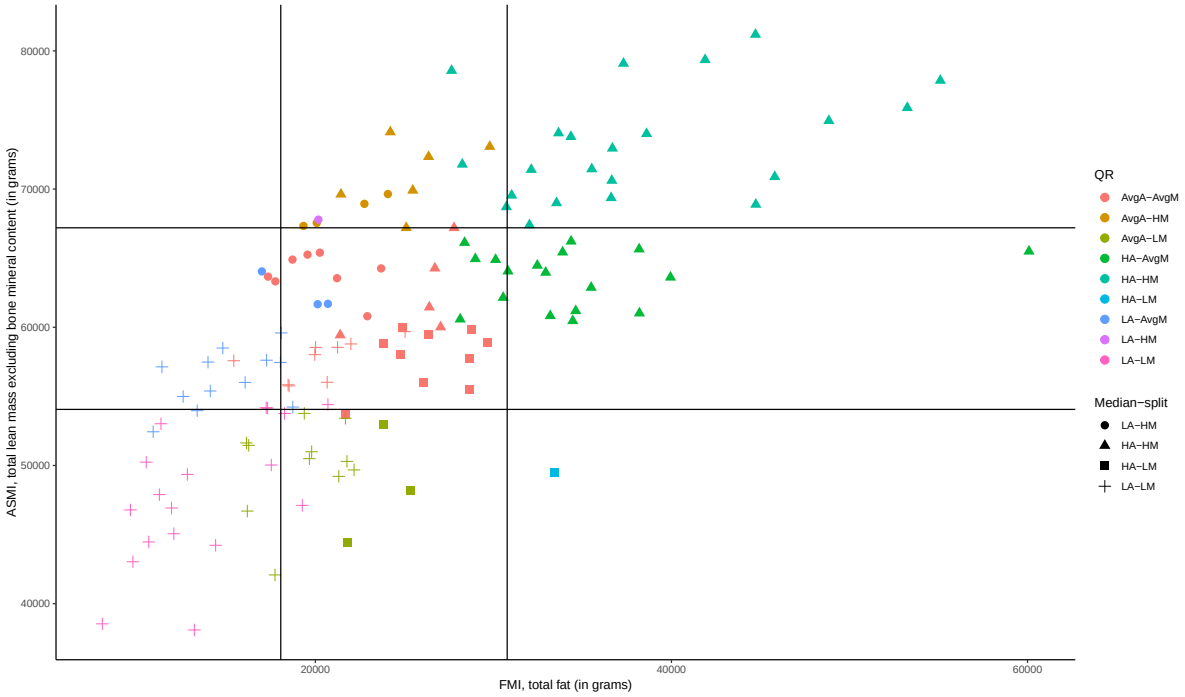
Notes: In-sample and out-of-sample profile analysis of the adiposity measures for the AvgA_{QR}-AvgM_{QR} phenotype, bucketed in terms of the median-split phenotypes.

Figure 22: QR and median-split phenotype clusters.

(a) NHANES 2011-2016.



(b) NHANES 2017-2018.



Notes: The vertical bars correspond to the 25% and 75% sample quantiles for the FMI index, whereas the horizontal bars correspond to the 25% and 75% sample quantiles for the ASMI index. For purposes of demonstration, we illustrate the overlap between the QR and median-split phenotypes for white males between the ages of 20 and 60 years, as this corresponds to our largest sub-group.

Table 16: Model comparison for health outcomes.

	QR	Median-split	BMI	WC	Total fat %
NHANES 2011-2016: AIC					
MetS	13613	13768	*12814	12976	13778
Depression	6120	6150	6124	*6115	6130
Short Sleep	8416	8407	8397	*8381	8400
General Health	9395	9496	*9336	9384	9456
IADL	2696	2685	*2664	2675	2681
Co-morbidity	8520	8533	8512	*8413	8476
NHANES 2017-2018: LR+					
MetS	2.19	1.82	*2.62	2.57	1.46
Depression	1.14	1.33	1.36	*1.45	1.15
Short Sleep	1.00	0.90	*1.09	0.98	0.97
General Health	1.28	1.52	*1.68	1.59	1.22
IADL	0.93	1.17	1.05	*1.20	1.10
Co-morbidity	*1.65	1.43	1.56	1.46	1.15
NHANES 2017-2018: LR-					
MetS	0.64	0.44	0.44	*0.19	0.21
Depression	0.90	0.67	0.77	*0.58	0.64
Short Sleep	1.00	1.11	*0.94	1.02	1.07
General Health	0.81	0.74	0.63	*0.53	0.53
IADL	1.04	0.90	0.89	0.76	*0.70
Co-morbidity	0.83	0.77	0.67	*0.59	0.64

Notes: The QR model reference group = AvgA-AvgM. The median-split model reference group = LA-HM. The BMI model reference group = normal. The WC model reference group = normal. The Total fat % model reference group = normal. The optimal values for each health outcome are displayed with an asterisk, that is lowest AIC, highest LR+, lowest LR-.

5 Conclusion

In chapter 2, we propose a method to forecast the distribution of electricity DART spreads, which are characterized by negative and positive spikes, seasonal behaviors, and complex dependencies on various covariates. A three-regime mixture model is developed and spike regimes are modeled using a Generalized Pareto Distribution. Both the frequency and severity components of the mixture model depend on covariates, whereby in-sample nested model comparisons reveal that their inclusion in the severity component is more critical. Using SAGE, we find that natural gas futures prices have the largest (relative) impact on model fit; moreover, regularization improves out-of-sample performance. In addition, the neural-network-based quantile regression benchmark fails to outperform the mixture model on the test set, which thus highlights that the improved interpretability of our model does not come at the expense of accuracy. Future areas of investigation include: (1) assessing the model's robustness to alternative data sets and (2) modeling the joint distribution between multiple zones, which can be used to construct locational differential trading strategies.

In chapter 3, we revisit whether energy futures can add value to an equity portfolio, taking into account the effects of the COVID-19 pandemic and the Russia-Ukraine war, which have profoundly impacted the energy sector. In the literature, many papers that study the effects of these crises focus on volatility and correlations, whereas we additionally assess the financial performance of equity portfolios enhanced with energy futures. Although many studies omit electricity as an investment opportunity due to its complex price dynamics, we include it alongside crude oil and natural gas. We therefore propose an empirically-based mean-variance asset allocation strategy to enhance the risk-adjusted return profile of an equity portfolio using energy futures. Our model can dynamically adjust to changing market conditions, and our results demonstrate an improvement in the out-of-sample Sharpe and Sortino ratios relative to the static S&P 500 benchmark portfolio, specifically during crisis periods. In addition, imposing moderate bounds on energy futures positions helps stabilize allocations and reduces risk. Despite the increased financialization of commodity markets, energy futures can improve the out-of-sample risk-return performance of a traditional equity portfolio. A contributing factor is the divergent behaviour of the various energy commodities in different periods. For example, electricity and natural gas, which are strongly correlated in regular time periods, show opposite trends during the Russia-Ukraine war. Future areas of investigation include: (1) using explanatory variables to improve forecasts of expected returns, volatilities, and correlations and (2) exploring additional energy commodities, such as wind and solar.

In chapter 4, we use quantile regression (QR) as an alternative to the LMS methodology to construct DEXA-derived phenotypes, as it is better suited to capture higher moments in the underlying data. The QR model generally shows a better (lower) AIC than the median-split model across health outcomes. Out-of-sample, the QR model has a better (higher) LR+ (positive likelihood ratio) than the median-split model for MetS and comorbidity. However, the QR model consistently underperforms in LR- (negative likelihood ratio) compared to the median-split model. Overall, the BMI (body mass index) and WC (waist circumference) models consistently demonstrate the best in-sample fit and out-of-sample performance among the models evaluated. Although QR may not outperform BMI and WC in general, it could potentially be better among certain subpopulations. This therefore needs to be explored more fully in future work. Whether the classification performances diverge in longitudinal studies should also be investigated.

This collection of manuscripts highlights the importance of understanding tail risks and extreme

events in a variety of domains, ranging from risk and portfolio management to epidemiology. Across these different contexts, we show the importance of accounting for nonlinear, non-Gaussian data that exhibit regime changes and structural breaks. Using flexible methodologies such as mixture models, dynamic asset allocation, and quantile regression, our studies demonstrate the value added from being able to model rare but critical events that impact real-world dynamics. Despite the different use cases considered, an overarching theme emerges: the ability to generate distributional forecasts and model extreme events, while accounting for potential structural breaks in the data, is necessary for informed decision making under uncertainty. Whether we are trying to hedge a financial position, improve the risk-adjusted performance of a portfolio, or develop targeted interventions in vulnerable sub-populations, we need to account for tail risk. Future work would benefit from the continued exploration of dynamic and interpretable statistical models to help guide critical decision making.

References

Abdussamad, A. M. and Inayat, A. (2024). Addressing limitations of the K-means clustering algorithm: Outliers, non-spherical data, and optimal cluster selection. *AIMS Math*, 9:25070–25097.

Adams, Z., Collot, S., and Kartsakli, M. (2020). Have commodities become a financial asset? Evidence from ten years of financialization. *Energy Economics*, 89:104769.

Adekoya, O. B. and Oliyide, J. A. (2021). How COVID-19 drives connectedness among commodity and financial markets: Evidence from TVP-VAR and causality-in-quantiles techniques. *Resources Policy*, 70:101898.

Afrasiabi, M., Aghaei, J., Afrasiabi, S., and Mohammadi, M. (2022). Probability density function forecasting of electricity price: Deep Gabor convolutional mixture network. *Electric Power Systems Research*, 213:108325.

Ashwell, M., Gunn, P., and Gibson, S. (2012). Waist-to-height ratio is a better screening tool than waist circumference and BMI for adult cardiometabolic risk factors: systematic review and meta-analysis. *Obesity Reviews*, 13(3):275–286.

Baeyens, J.-P., Serrien, B., Goossens, M., and Clijsen, R. (2019). Questioning the "SPIN and SNOUT" rule in clinical testing. *Archives of Physiotherapy*, 9:1–6.

Baker, S. D. (2021). The financialization of storable commodities. *Management Science*, 67(1):471–499.

Basak, S. and Pavlova, A. (2016). A model of financialization of commodities. *The Journal of Finance*, 71(4):1511–1556.

Beccuti, G. and Pannain, S. (2011). Sleep and obesity. *Current Opinion in Clinical Nutrition & Metabolic Care*, 14(4):402–412.

Belousova, J. and Dorfleitner, G. (2012). On the diversification benefits of commodities from the perspective of Euro investors. *Journal of Banking & Finance*, 36(9):2455–2472.

Benth, F. E., Kallsen, J., and Meyer-Brandis, T. (2007). A non-Gaussian Ornstein–Uhlenbeck process for electricity spot price modeling and derivatives pricing. *Applied Mathematical Finance*, 14(2):153–169.

Bessler, W. and Wolff, D. (2015). Do commodities add value in multi-asset portfolios? An out-of-sample analysis for different investment strategies. *Journal of Banking & Finance*, 60:1–20.

Bhardwaj, G., Gorton, G., and Rouwenhorst, G. (2015). Facts and fantasies about commodity futures ten years later. Technical report, National Bureau of Economic Research.

Bishop, C. M. (1994). Mixture density networks. Neural Computing Research Group Report: NCRG/94/004, Aston University. Available at `https://publications.aston.ac.uk/id/eprint/373/1/NCRG_94_004.pdf`.

Borgards, O., Czudaj, R. L., and Van Hoang, T. H. (2021). Price overreactions in the commodity futures market: An intraday analysis of the COVID-19 pandemic impact. *Resources Policy*, 71:101966.

Brusaferri, A., Matteucci, M., Ramin, D., Spinelli, S., and Vitali, A. (2020). Probabilistic day-ahead energy price forecast by a mixture density recurrent neural network. In *2020 7th International Conference on Control, Decision and Information Technologies (CoDIT)*, pages 523–528.

Carmona, R. (2015). Financialization of the commodities markets: A non-technical introduction. *Commodities, Energy and Environmental Finance*, pages 3–37.

CDC and NCHS (2021). Technical Documentation for the 1999-2004 Dual-Energy X-Ray Absorptiometry (DXA) Multiple Imputation Data Files.

Cheng, I.-H. and Xiong, W. (2014). Financialization of commodity markets. *Annual Reviews of Financial Economics*, 6(1):419–441.

Cheung, C. S. and Miu, P. (2010). Diversification benefits of commodity futures. *Journal of International Financial Markets, Institutions and Money*, 20(5):451–474.

Chishti, M. Z., Khalid, A. A., and Sana, M. (2023). Conflict vs sustainability of global energy, agricultural and metal markets: A lesson from Ukraine-Russia war. *Resources Policy*, 84:103775.

Christoffersen, P. and Pan, X. N. (2018). Oil volatility risk and expected stock returns. *Journal of Banking & Finance*, 95:5–26.

Cole, T. J. and Green, P. J. (1992). Smoothing reference centile curves: the lms method and penalized likelihood. *Statistics in Medicine*, 11(10):1305–1319.

Covert, I., Lundberg, S. M., and Lee, S.-I. (2020). Understanding global feature contributions with additive importance measures. *Advances in Neural Information Processing Systems*, 33:17212–17223.

Creti, A., Joëts, M., and Mignon, V. (2013). On the links between stock and commodity markets' volatility. *Energy Economics*, 37:16–28.

Curtin, L. R., Mohadjer, L. K., Dohrmann, S. M., Montaquila, J. M., Kruszan-Moran, D., Mirel, L. B., Carroll, M. D., Hirsch, R., Schober, S., and Johnson, C. L. (2012). The National Health and Nutrition Examination Survey: Sample Design, 1999-2006. *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, (155):1–39.

Das, R., Bo, R., Chen, H., Rehman, W. U., and Wunsch, D. (2022). Forecasting nodal price difference between day-ahead and real-time electricity markets using long-short term memory and sequence-to-sequence networks. *IEEE Access*, 10:832–843.

Daskalaki, C. and Skiadopoulos, G. (2011). Should investors include commodities in their portfolios after all? New evidence. *Journal of Banking & Finance*, 35(10):2606–2626.

Davison, M., Anderson, C. L., Marcus, B., and Anderson, K. (2002). Development of a hybrid model for electrical power spot prices. *IEEE Transactions on Power Systems*, 17(2):257–264.

Degiannakis, S., Filis, G., and Arora, V. (2018). Oil prices and stock markets: A review of the theory and empirical evidence. *The Energy Journal*, 39(5):85–130.

Demidova-Menzel, N. and Heidorn, T. (2007). Commodities in asset management. Technical report, Frankfurt School-Working Paper Series.

Demiralay, S., Bayraci, S., and Gaye Gencer, H. (2019). Time-varying diversification benefits of commodity futures. *Empirical Economics*, 56:1823–1853.

Deschatre, T., Féron, O., and Gruet, P. (2021). A survey of electricity spot and futures price models for risk management applications. *Energy Economics*, 102:105504.

Després, J.-P. (2012). Body fat distribution and risk of cardiovascular disease: An update. *Circulation*, 126(10):1301–1313.

Dixon, J. B. (2010). The effect of obesity on health outcomes. *Molecular and Cellular Endocrinology*, 316(2):104–108.

Du, S. (2005). Commodity futures in asset allocation.

Elsayed, A. H., Nasreen, S., and Tiwari, A. K. (2020). Time-varying co-movements between energy market and global financial markets: Implication for portfolio diversification and hedging strategies. *Energy Economics*, 90:104847.

Flegal, K. M. and Cole, T. J. (2013). *Construction of LMS parameters for the Centers for Disease Control and Prevention 2000 growth charts*. Number 63. US Department of Health and Human Services, Centers for Disease Control and Prevention.

Fox, E., Simons, R., Moskatov, O., and Russo, M. (2019). New York ISO climate load impact study preliminary forecasts. Technical report, Itron. Available at `https://www.nyiso.com/documents/20142/8115627/ClimateStudy.pdf`.

Frankenfield, D. C., Rowe, W. A., Cooney, R. N., Smith, J. S., and Becker, D. (2001). Limits of body mass index to detect obesity and predict body composition. *Nutrition*, 17(1):26–30.

Gaete, M. and Herrera, R. (2023). Diversification benefits of commodities in portfolio allocation: A dynamic factor copula approach. *Journal of Commodity Markets*, 32:100363.

Galarneau-Vincent, R., Gauthier, G., and Godin, F. (2023). Foreseeing the worst: Forecasting electricity DART spikes. *Energy Economics*, 119:106521.

Gao, L., Hitzemann, S., Shaliastovich, I., and Xu, L. (2022). Oil volatility risk. *Journal of Financial Economics*, 144(2):456–491.

Geman, H. (2005). *Commodities and commodity derivatives: Modeling and pricing for agriculturals, metals and energy*. John Wiley & Sons.

Giamouridis, D., Sakkas, A., and Tessaromatis, N. (2014). The role of commodities in strategic asset allocation. In *27th Australasian Finance and Banking Conference*.

Gorton, G. and Rouwenhorst, K. G. (2006). Facts and fantasies about commodity futures. *Financial Analysts Journal*, 62(2):47–68.

Hansen-Tangen, N. H. and Overaae, M. (2015). Do commodities offer diversification benefits in multi-asset portfolios? Master's thesis, Norwegian University of Life Sciences, Ås.

Hattori, K., Tatsumi, N., and Tanaka, S. (1997). Assessment of body composition by using a new chart method. *American Journal of Human Biology: The Official Journal of the Human Biology Association*, 9(5):573–578.

Henderson, B. J., Pearson, N. D., and Wang, L. (2015). New evidence on the financialization of commodity markets. *The Review of Financial Studies*, 28(5):1285–1311.

Heymsfield, S. B., Smith, R., Aulet, M., Bensen, B., Lichtman, S., Wang, J., and Pierson Jr, R. (1990). Appendicular skeletal muscle mass: measurement by dual-photon absorptiometry. *The American Journal of Clinical Nutrition*, 52(2):214–218.

Huber, P. J. and Ronchetti, E. M. (2011). *Robust statistics*. John Wiley & Sons.

Jaeschke, R., Guyatt, G. H., Sackett, D. L., Guyatt, G., Bass, E., Brill-Edwards, P., Browman, G., Cook, D., Farkouh, M., Gerstein, H., et al. (1994). Users' guides to the medical literature: Iii. How to use an article about a diagnostic test b. What are the results and will they help me in caring for my patients? *Jama*, 271(9):703–707.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, pages 163–172.

Jensen, G. R., Johnson, R. R., and Mercer, J. M. (2002). Tactical asset allocation and commodity futures. *Journal of Portfolio Management*, 28(4):100.

Jindai, K., Nielson, C. M., Vorderstrasse, B. A., and Quiñones, A. R. (2016). Peer reviewed: multimorbidity and functional limitations among adults 65 or older, NHANES 2005–2012. *Preventing Chronic Disease*, 13.

Kakinami, L., Plummer, S., Cohen, T. R., Santosa, S., and Murphy, J. (2022). Body-composition phenotypes and their associations with cardiometabolic risks and health behaviours in a representative general US sample. *Preventive Medicine*, 164:107282.

Kang, W., Tang, K., and Wang, N. (2023). Financialization of commodity markets ten years later. *Journal of Commodity Markets*, 30:100313.

King, D. E., Xiang, J., and Pilkerton, C. S. (2018). Multimorbidity trends in United States adults, 1988–2014. *The Journal of the American Board of Family Medicine*, 31(4):503–513.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Klüppelberg, C., Meyer-Brandis, T., and Schmidt, A. (2010). Electricity spot price modelling with a view towards extreme spike risk. *Quantitative Finance*, 10(9):963–974.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156.

Kroenke, K., Spitzer, R. L., and Williams, J. B. (2001). The PHQ-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

Lean, H. H., Nguyen, D. K., Sensoy, A., and Uddin, G. S. (2023). On the role of commodity futures in portfolio diversification. *International Transactions in Operational Research*, 30(5):2374–2394.

Liu, Q. and Tu, A. H. (2012). Jump spillovers in energy futures markets: Implications for diversification benefits. *Energy Economics*, 34(5):1447–1464.

Longstaff, F. A. and Wang, A. W. (2004). Electricity forward prices: A high-frequency empirical analysis. *The Journal of Finance*, 59(4):1877–1900.

Lukaski, H. C. (1987). Methods for the assessment of human body composition: Traditional and new. *The American Journal of Clinical Nutrition*, 46(4):537–556.

Marcjasz, G., Narajewski, M., Weron, R., and Ziel, F. (2023). Distributional neural networks for electricity price forecasting. *Energy Economics*, 125:106843.

Miller, B., Fridline, M., Liu, P.-Y., and Marino, D. (2014). Use of CHAID decision trees to formulate pathways for the early detection of metabolic syndrome in young adults. *Computational and Mathematical Methods in Medicine*, 2014(1):242717.

Minetto, M. A., Busso, C., Lalli, P., Gamerro, G., and Massazza, G. (2021). DXA-derived adiposity and lean indices for management of cardiometabolic and musculoskeletal frailty: data interpretation tricks and reporting tips. *Frontiers in Rehabilitation Sciences*, 2:712977.

Narayan, P. K. and Gupta, R. (2015). Has oil price predicted stock returns for over a century? *Energy Economics*, 48:18–23.

Nguyen, D. K., Topaloglou, N., and Walther, T. (2020). Asset classes and portfolio diversification: Evidence from a stochastic spanning approach. *Available at SSRN 3721907*.

NHANES (2016). MEC laboratory procedures manual. `https://wwwn.cdc.gov/nchs/data/nhanes/public/2015manuals/2016_MEC_Laboratory_Procedures_Manual.pdf`. [Online; accessed 15-January-2025].

NHANES (2024). Tutorials. `https://wwwn.cdc.gov/nchs/nhanes/tutorials/default.aspx`. [Online; accessed 8-July-2024].

Noguera-Santaella, J. (2016). Geopolitics and the oil price. *Economic Modelling*, 52:301–309.

Nowak-Brzezińska, A. and Gaibei, I. (2022). How the outliers influence the quality of clustering? *Entropy*, 24(7):917.

Nuttall, F. Q. (2015). Body mass index: obesity, BMI, and health: a critical review. *Nutrition Today*, 50(3):117–128.

NYISO (2021). Load forecast uncertainty modeling: New York temperature distribution analysis. Technical report, New York Independent System Operator. Available at https://www.nyiso.com/documents/20142/20053386/LFU_Whitepaper_Phase1_Results.pdf.

Ode, J. J., Pivarnik, J. M., Reeves, M. J., and Knous, J. L. (2007). Body mass index as a predictor of percent fat in college athletes and nonathletes. *Medicine and Science in Sports and Exercise*, 39(3):403–409.

Pickands III, J. (1975). Statistical inference using extreme order statistics. *The Annals of Statistics*, 3(1):119–131.

Prado, C. M., Siervo, M., Mire, E., Heymsfield, S. B., Stephan, B. C., Broyles, S., Smith, S. R., Wells, J. C., and Katzmarzyk, P. T. (2014). A population-based approach to define body-composition phenotypes. *The American Journal of Clinical Nutrition*, 99(6):1369–1377.

Reichmann, W. M., Katz, J. N., and Losina, E. (2011). Differences in self-reported health in the osteoarthritis initiative (OAI) and third national health and nutrition examination survey (NHANES-III). *PLoS One*, 6(2):e17345.

Rothman, K. J. (2008). BMI-related errors in the measurement of obesity. *International Journal of Obesity*, 32(3):S56–S59.

Rousseeuw, P. J. and Hubert, M. (2011). Robust statistics for outlier detection. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1):73–79.

Silvennoinen, A. and Thorp, S. (2013). Financialization, crisis and commodity correlation dynamics. *Journal of International Financial Markets, Institutions and Money*, 24:42–65.

Šimundić, A.-M. (2009). Measures of diagnostic accuracy: basic definitions. *EJIFCC*, 19(4):203.

Tang, K. and Xiong, W. (2012). Index investment and the financialization of commodities. *Financial Analysts Journal*, 68(6):54–74.

Tewari, N., Awad, S., Macdonald, I. A., and Lobo, D. N. (2018). A comparison of three methods to assess body composition. *Nutrition*, 47:1–5.

Thomas, D. M., Crofford, I., Scudder, J., Oletti, B., Deb, A., and Heymsfield, S. B. (2025). Updates on methods for body composition analysis: Implications for clinical practice. *Current Obesity Reports*, 14(1):1–20.

Wannamethee, S. G. and Atkins, J. L. (2015). Muscle loss and obesity: The health implications of sarcopenia and sarcopenic obesity. *Proceedings of the Nutrition Society*, 74(4):405–412.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, 30(4):1030–1081.

Woolcott, O. O. and Bergman, R. N. (2018). Relative fat mass (rfm) as a new estimator of whole-body fat percentage – A cross-sectional study in American adult individuals. *Scientific Reports*, 8(1):10980.

Zakeri, B., Paulavets, K., Barreto-Gomez, L., Echeverri, L. G., Pachauri, S., Boza-Kiss, B., Zimm, C., Rogelj, J., Creutzig, F., Ürge-Vorsatz, D., et al. (2022). Pandemic, war, and global energy transitions. *Energies*, 15(17):6114.

Zamboni, M., Mazzali, G., Fantin, F., Rossi, A., and Di Francesco, V. (2008). Sarcopenic obesity: A new category of obesity in the elderly. *Nutrition, Metabolism and Cardiovascular Diseases*, 18(5):388–395.

Zhang, Q., Hu, Y., Jiao, J., and Wang, S. (2024). The impact of Russia–Ukraine war on crude oil prices: An EMC framework. *Humanities and Social Sciences Communications*, 11(1):1–12.

A Continuous Ranked Probability Score (CRPS)

Let X be a random variable, with F being the CDF of its predicted distribution. Assume x is a realization (i.e. observation) of X. The CRPS for such observation is defined as

$$\mathrm{CRPS}(F,x) = \int_{-\infty}^{\infty} (F(y) - \mathscr{H}(y-x))^2 \, dy$$

where $\mathscr{H}$ denotes the Heaviside step function: $\mathscr{H}(z) = \mathbb{1}_{\{z \geq 0\}}$. It measures the concentration of the predicted distribution around the realized value. The smallest possible value of the CRPS is 0, which occurs if the predictive distribution is degenerate and fully concentrated on x. Conversely, the larger the CRPS is, the less concentrated around x the predicted distribution will be.

The total CRPS score for a predictive model is obtained by averaging CRPS scores over all observations:

$$\mathrm{CRPS} = \frac{1}{\tau} \sum_{t=1}^{\tau} \mathrm{CRPS}(F_t, x_t),$$

where x_t is observation t, F_t is its predicted distribution, and τ denotes the total number of observations. The CRPS generalizes the MAE and reduces to the MAE if the forecasts are deterministic, i.e., if predictive distributions F_t are degenerate for all t.

B Hidden Markov Model (HMM)

Let $\{\nu_n\}_{n \in \mathbb{N}}$ represent a discrete Markov chain with latent states indexed by $j \in \{1, \ldots, K\}$. The probability density function (PDF) of the excess return $E_t^{(X)}$, where $X \in \{S, F^{(E)}, F^{(C)}, F^{(N)}\}$, is denoted by

$$f^{(j)}(e_t^{(X)}) = f_{E_t^{(X)} | \nu_t, \nu_{1:t-1}, E_{1:t-1}^{(X)}} \left(e_t^{(X)} \mid j, \nu_{1:t-1}, E_{1:t-1}^{(X)} \right),$$

where $f^{(1)}, \ldots, f^{(K)}$ constitutes a set of densities corresponding to the latent states. We define the predictive probabilities at time t as

$$\eta_{t,i} = \mathbb{P}\left(\nu_t = i \mid E_{1:t-1}^{(X)} = e_{1:t-1}^{(X)} \right),$$

where $\eta_{t,i}$ denotes the probability that the regime at time t is i, conditional on the information (filtration) available up to time $t-1$. The predictive probabilities $\eta := \{(\eta_{t,1}, \cdots, \eta_{t,K})\}_{t=1}^{n}$ can be computed recursively using

$$\eta_{t+1,i} = \frac{\sum_{j=1}^{K} P_{j,i} f^{(j)}(e_t^{(X)}) \eta_{t,j}}{\sum_{l=1}^{K} f^{(l)}(e_t^{(X)}) \eta_{t,l}},$$

where $P_{j,i} = \mathbb{P}(\nu_{t+1} = i \mid \nu_t = j)$ is the (j,i)-th entry of the transition matrix, that is, the probability to transition from state j to i. The conditional density is given by

$$f_{E_t^{(X)} | E_1^{(X)}, \ldots, E_{t-1}^{(X)}}(e_t^{(X)} \mid e_1^{(X)}, \ldots, e_{t-1}^{(X)}) = \sum_{j=1}^{K} f^{(j)}(e_t^{(X)}) \, \eta_{t,j},$$

and so the maximum likelihood estimate of the set of parameters is found using

$$\hat{\Theta} = \arg\max_{\Theta} \sum_{t=1}^{n} \log \left(\sum_{j=1}^{K} f_{\Theta}^{(j)}(e_t^{(X)}) \eta_{t,j} \right),$$

where we assume that $\eta_{1,j} = \pi_j$ for $j = 1, \cdots, K$, that is, the initial distribution of the predictive probabilities is set equal to the stationary probabilities. In the case of an HMM with two hidden states and with Gaussian density functions, we have

$$\Theta = [\mu_1, \mu_2, \sigma_1, \sigma_2, P_{1,1}, P_{2,2}],$$

and if we replace the Gaussian densities with NIG densities, we get

$$\Theta^{\text{NIG}} = \left[\alpha_1, \alpha_2, \beta_1, \beta_2, \delta_1, \delta_2, \mu_1^{\text{NIG}}, \mu_2^{\text{NIG}}, P_{1,1}^{\text{NIG}}, P_{2,2}^{\text{NIG}} \right].$$

Moreover, we have that

$$f_{\mathcal{N}} = \mathcal{N}(\mu, \sigma^2),$$

$$f_{\text{NIG}} = \frac{\alpha\delta \exp(\delta\gamma) K_1 \left(\alpha\sqrt{\delta^2 + (x-\mu)^2} \right) \exp\left(\beta(x-\mu)\right)}{\pi\sqrt{\delta^2 + (x-\mu)^2}}.$$

The most likely sequence of latent states, given a sequence of observations, is then computed using the Viterbi algorithm. That is, first, calculate

$$\delta_j(1) = f_{\nu_1}(j) f_{E_1^{(X)}|\nu_1}(e_1^{(X)} \mid j), \quad j = 1, \ldots, K.$$

Then, for $t = 1, \ldots, n-1$, calculate

$$\Psi_j(t) = \arg\max_{k \in \{1,\ldots,K\}} \left[\delta_k(t) P_{k,j} \right], \quad j = 1, \ldots, K,$$

$$\delta_j(t+1) = \max_{k \in \{1,\ldots,K\}} \left[\delta_k(t) P_{k,j} \right] f_{E_{t+1}^{(X)}|\nu_{t+1}}(e_{t+1}^{(X)} \mid j), \quad j = 1, \ldots, K,$$

such that $\Psi_j(t)$ records which state k leads to state j at time $t+1$ with the highest probability path, and where $\delta_j(t+1)$ is the probability of the most likely path that ends in state j at time $t+1$. Finally, proceed recursively to get the most likely path. First,

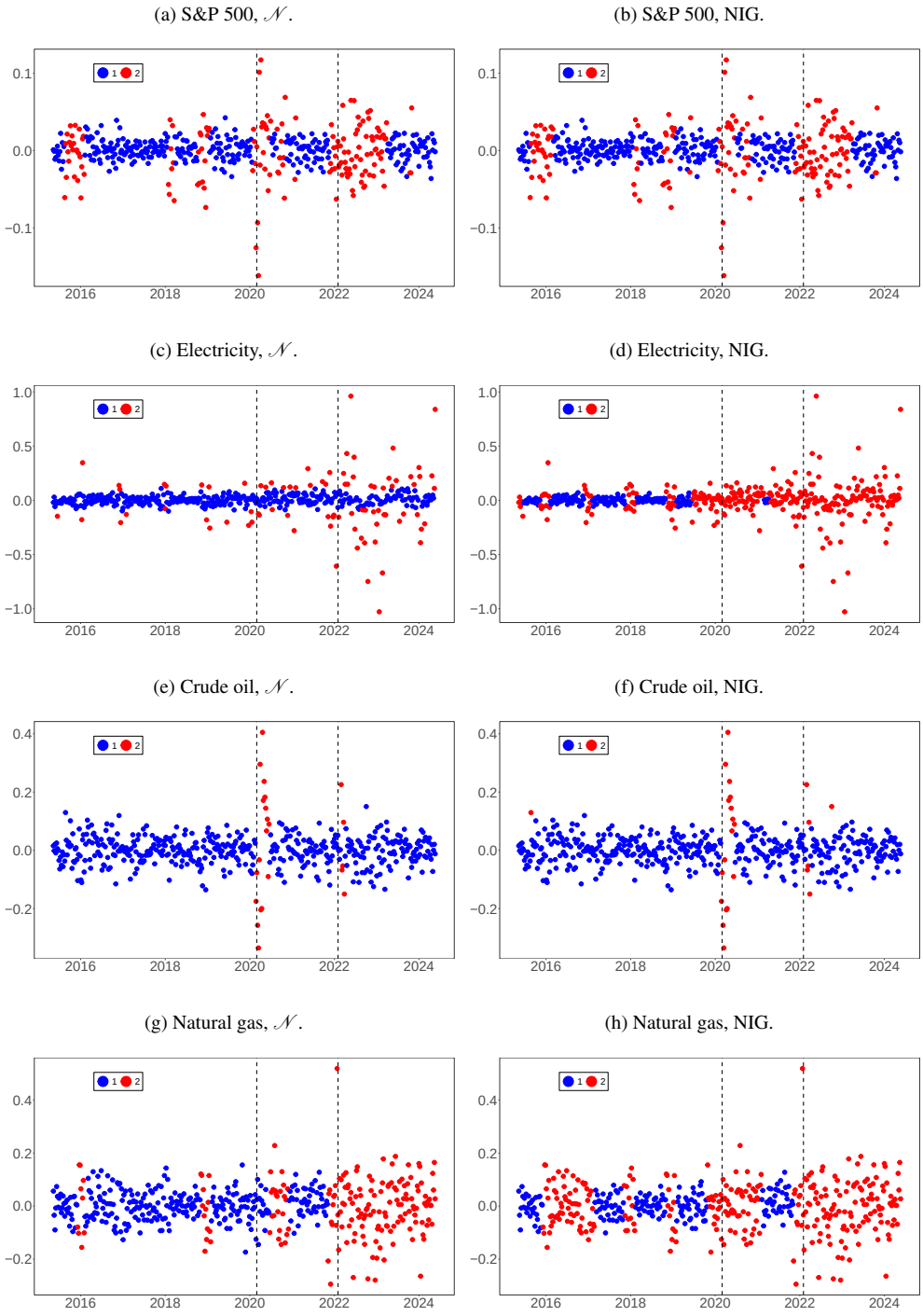
$$\nu_n^* = \arg\max_{k \in \{1,\ldots,K\}} \delta_k(n),$$

which gives the final state of the most likely path. Then, for $t = n-1, \ldots, 1$,

$$\nu_t^* := \Psi_{\nu_{t+1}^*}(t),$$

where we go backwards in time to recover the optimal sequence of latent states. In Figure 23 below, we present the most likely paths for each asset, using an HMM with two hidden states, for both Gaussian and NIG densities.

Figure 23: HMM latent states of weekly excess returns.



Notes: The data consists of weekly excess returns, ranging between May 12, 2014 to May 21, 2024. The vertical bars demarcate the different periods considered. That is, the normal period ranges between May 12, 2014 to March 12, 2020, the COVID-19 period ranges between March 12, 2020 to February 24, 2022, and the Russia-Ukraine period ranges between February 24, 2022 to May 21, 2024.

C LMS methodology

The LMS parameters consist of the power in the Box-Cox transformation (L), the median (M), and the generalized coefficient of variation (S) (Flegal and Cole (2013)). Assuming normality, if we let A denote the value of the anthropometric variable and let Z refer to the desired percentile in standard deviation units, then we have that

$$A = \begin{cases} M(1+LSZ)^{1/L}, & \text{if } L \neq 0 \\ M\exp(SZ), & \text{if } L = 0 \end{cases}$$

or, conversely, given a value of A, the corresponding z-score Z is given by

$$Z = \begin{cases} [(X/M)^L - 1]/LS, & \text{if } L \neq 0 \\ \ln(X/M)/S, & \text{if } L = 0 \end{cases}$$

where the method of maximum penalized likelihood is used to obtain values for L, M, and S, smoothed with respect to either age or height (Cole and Green (1992)). The LMS methodology captures the mean, variance, and skewness of a distribution.