# Multivariate Scenario Generation – An Arima and Copula Approach

S. Mishra, C. Würsig, C. Bordin, and I. Palu

***Abstract*—In mathematical optimization uncertainty is expressed through scenarios. auto-regressive integrated moving average (ARIMA) is one of the known practice to generate scenarios. This paper is about scenario generation using multivariate data: electrical power demand, wind power generation and energy market price. An ARIMA model along with Copula is implemented for scenario generation. The results are presented and discussed.**

***Index Terms*—Multivariate scenario generation, ARIMA, Copula, Stochastic programming**

## I. Introduction

This work is about scenario generation using copula. Scenario generation is an important part of stochastic programming. The generated scenarios however should retain the original statistical properties of the data. Auto regressive inter grated moving average (ARIMA) has been used extensively in literature [1] to generate scenarios. One of the drawbacks of ARIMA is the applicability to multivariate distributions. To overcome this Copula is used to generate scenarios as presented here [2], [3]. A regular vine copula and the goodness of fit measures are discussed here [4]. A Bayes theory-based copula is presented here. [5]. A comprehensive study of various copula models with real world data is presented here [6]. A multivariate copula-based forecasting method is explained here [7].

Multivariate copula is gaining more importance due to the nature and availability of data and relations among them. In this paper wind, demand and price data are considered as the multivariate data. Copula is used to generate multivariate distributions. These are sampled using ARIMA and the results are presented. The rest of the paper is organized in four sections: scenario generation, computational experiments, discussion and conclusion.

## II. Scenario Generation

This section describes the mathematical model for the scenario generation using copula. This section is further divided into two subsections: ARIMA model and copula. The former presents a multivariate ARIMA formulation considering three variables. The later states the copula to sample the residuals.

Manuscript received October 1, 2018; revised April 1, 2019. This work was supported in part by the European Unison social development funds.

S. Mishra and I. Palu are with Tallinn University of Technology, Tallinn, Estonia (e-mail: sambeet.mishra@ttu.ee).

C. Würsig is with Leibniz Universität Hannover, Germany

C. Bordin is with SINTEF Energy Research, Trondheim, Norway

### A. ARIMA Model

The ARIMA Model is a widely used model for modeling [8]. We use the model, to capture the time series behavior of the series.

The authors consider the statistically correlated scenarios because the stochastic variables: wind, demand, and price are co-related. Thus, formulating the ARIMA $(\phi, \varphi)$ a quasi-contemporaneous stochastic process price $\left(y_{s,t}^{a}\right)$ and wind $\left(y_{s,t}^{c}\right)$ as in 1(a-c). The residuals $\varepsilon_{t,s}^{a}, \varepsilon_{t,s}^{b}, \varepsilon_{t,s}^{c}$ are statistically dependent. Thus, the dependency structure of the stochastic processes can be stated as $\varepsilon\{\varepsilon_{t,s}^{a} \cdot \varepsilon_{t-j,s}^{b} \cdot \varepsilon_{t-j,s}^{c}\} \neq 0$. $\varepsilon_{s}^{a}, \varepsilon_{s}^{b}, \varepsilon_{s}^{c}$ are the series of errors simulated to produce residual cross-correlogram of stochastic process. In 1(d) the error correlation between stochastic process a & b, a & c are presented and finally reduced to a product of an orthogonal matrix B and identity matrix $\psi(E[\psi \cdot \psi^{T}] = I)$. The cross correlation between $\varepsilon_{t,s}^{a}$ and $\varepsilon_{t,s}^{b}$ can be represented through variance-covariance matrix G. G is essentially a positive semi-definite and symmetric matrix. This matrix is further decomposed using Cholesky decomposition $(G = LL^{T})$ [9]-[11]. L is the upper triangular matrix that is also the orthogonal matrix $(B = L)$.

$$y_{s,t}^{a} = \sum_{j=1}^{\eta^{a}} \phi_{j}^{a} \cdot y_{i-j,s}^{a} + \varepsilon_{s,t}^{a} - \sum_{j=1}^{\tau^{a}} \varphi_{j}^{a} \cdot \varepsilon_{t-j,s}^{a} \quad (1a)$$

$$\text{v}\ y_{s,t}^{b} = \sum_{j=1}^{\eta b} \phi_{j}^{b} \cdot y_{i-j,s}^{b} + \varepsilon_{s,t}^{b} - \sum_{j=1}^{\tau b} \varphi_{j}^{b} \cdot \varepsilon_{t-j,s}^{b} \quad (1b)$$

$$y_{s,t}^{c} = \sum_{j=1}^{\eta c} \phi_{j}^{c} \cdot y_{i-j,s}^{c} + \varepsilon_{s,t}^{c} - \sum_{j=1}^{\tau c} \varphi_{j}^{c} \cdot \varepsilon_{t-j,s}^{c} \quad (1c)$$

$$\varepsilon_{s,t}^{1} = \begin{pmatrix} \varepsilon_{s,t}^{a} \\ \varepsilon_{s,t}^{b} \end{pmatrix} \quad \varepsilon_{s,t}^{2} = \begin{pmatrix} \varepsilon_{s,t}^{a} \\ \varepsilon_{s,t}^{c} \end{pmatrix} \Rightarrow \varepsilon = \begin{pmatrix} \varepsilon_{s,t}^{1} \\ \varepsilon_{s,t}^{2} \end{pmatrix} \Rightarrow \varepsilon = B\psi \quad (1d)$$

$$G = \text{cov}(\varepsilon, \varepsilon^{T}) = BB^{T} \quad (1e)$$

### B. Copula

The residuals of the ARIMA Model are fitted to a Copula Model to capture time varying dependence of the data. The authors use for this purpose R-vine copulas introduced by Bedford and Cooke (2001b, 2002).

The general theory for copulas is Skalars Theorem (1959), based on this Theorem, Skalar shows that a every multivariate distribution can be written as a multivariate copula function. Equation (2) shows Skalars Theorem applied to a three-dimensional dataset.

Following Skalar (1959) this density function is uniquely

DOI: 10.7763/IJMO.2019.V9.700

represented by the following form, if it is continuous.

$$F(a,b,c) = C\big(F_a(a), F_b(b), F_c(c)\big) \tag{2}$$

Joe (1996) makes this theorem usable for Vine Copulas, since he showed that Skalars Theorem can be decomposed to bivariate copulas. For a multivariate distribution with three variables it thus follows that this decomposition can uniquely identify the density function.

$$f(a|b,c) = c_{ac|b}\left(F_{a|b}(a|b), F_{c|b}(c|b)\right) f(a|b) \tag{3}$$

where

$$f(a|b) = c_{ab}\big(F_a(a), F_b(b)\big) f_a(a)$$

The R-vine (regular vine) model is chosen to model the multivariate dependence in this empirical application.

Fitting multivariate data to a copula is a challenging task, since commonly used copula models, like the normal copula, the t copula or the gumbel copula are either symmetric or have only one parameter to estimate the entire copula, which decreases the flexibility of the distribution. Bivariate copulas have a wider variety of choices, thus Kurowicka and Cooke (2006) developed the R-vine copula models that fit multiple bivariate copulas to the multivariate dataset and are thus able to capture the dependence structure of the multivariate dataset. The modeling scheme is based on a decomposition of a multivariate density into a cascade of pair copula (Aas et al. p.1). R-vine's are represented by a hierarchical tree structure, where the first tree is estimated by n-1 bivariate copula and the second by n-2 conditional on a single variable. For a three-dimensional dataset two copula need to be estimated directly and one conditional copula. To estimate the R-vine, Dissmann *et al.* (2012) developed a sequential search approach, they first estimate the family and parameters of the first tree via the AIC criterion. Then they use this result to estimate the second tree. Additionally, they employ a maximum spanning tree algorithm to choose an appropriate edge weight. This paper implements their method and estimation technique, in order to take advantage of the benefits of the diversity of bivariate copula.

### C. ARIMA Forecasting Using Copula

The approach used in this paper is reminiscent of the Copula GARCH model, introduced by Jondenau and Rockinger (2006). First the ARIMA model is estimated, with the standardized residuals of the ARIMA model the R-vine copula model is estimated. The R-vine Model is then estimated using the remaining errors terms from the ARIMA model to capture dependencies between the variables that the time series model ARIMA cannot capture. The Copula model is fitted to uniform [0,1] margins. Afterwards following Dissmann *et al*. (2012) we simulate from the copula model and transform the thereby obtained data using the not standardized residuals from the empirical ARIMA model as an empirical density function. To model the time series behavior, the simulation result is obtained using the sampled residuals and the fitted ARIMA model.

## III. Implementation

In the following we present the implementation of our method, this simulation is conducted on the logarithm of wind, price and demand variables for 100 times. The scripts are written in *R* statistics programming language.

We estimate missing data, via linear interpolation for single missing values. For wind we estimate the last month via an ARIMA forecast due to the unaccounted data for December. The ARIMA model is fitted based on the conditional sum of squares to find the starting values. Following that maximum likelihood to find the optimal parameter estimates with respect to the AIC criterion.

We use the residuals and standardize them to fit an R-Vine Copula onto the residuals. The tree structure is determined via pair-copula families and estimated sequentially. For the model families the AIC criterion is used, parameter values are estimated using maximum likelihood estimation.

Following [12] methods we simulate the uniform estimates from the R-Vine Copula model.

We transform the uniform values using the trimmed empirical quantile distribution of our residuals into simulated observations.

We enter the simulated estimates into the ARIMA model and obtain the results after taking the exponential function of the values.

## IV. Computational Experiments

The provided sample is hourly data for the year 2017, with the Price in \euro/MWh, Wind in MWH and Demand in MWH. The data contains two missing observations, they are interpolated, additionally the last 263 observations for Wind data are missing, in order to model this data an ARIMA model is fitted on the observed sample and the 263 missing values are estimated. The approach used is close to the GARCH Copula estimation, in place for a ARMA (p,q)-GARCH model and ARIMA model is used, since the data is unlikely heteroskedastic and it is unnecessary to model GARCH effects for this time series. This method enables us to fit the a copula approach easily to the data and to model the time series behavior.

First the data is fitted to an ARIMA model, that is optimally chosen based on the AIC criterion. The ARIMA process is required to be stationary and seasonal, this is necessary because of the limited amount of data, we are forecasting a year using only a year of data, trends cannot be captured reliably. It might be a substantial increase in wind production, but it is not clear if it is due to a windy year or additional wind farms, that would increase next year's production as well. The seasonality is assumed because of the nature of the data, wind is seasonal, as well as the demand, the price is seasonal as well. To ensure positivity of the data, we are fitting the natural logarithm of the data and transform them for analysis later. To minimize extreme observations in our data set, considering the large time frame we are trying to model, we trim the residuals at 3% (we remove the 3% lowest and the 3% highest values). With this value we have a near normal kurtosis, before the kurtosis for the price and the wind reached over 40. To ensure that our results remain robust for different cutoff values, we used multiple values, the results

are not inconsistent, the variation of the data increases as expected.

The estimated coefficients of the ARIMA model are presented in Table I, the standard errors for the coefficients are low and the model fit seems to be reasonable. In order to model serial dependence, the innovations need to be modeled, in order to model them we are standardizing the residuals and transforming them into uniform [0,1] margins. The best R-Vine copula model is chosen by optimizing the bi-variate copula models and choosing the best fit with the AIC criterion. We sample the residuals from the trimmed series, we draw them based on their assigned uniform [0,1] margins provided by the random sampling from the copula. In the next step we find the best R-Vine Copula using maximum likelihood estimation and the AIC criterion. Simulations are conducted from this R-Vine structure. The result are uniform [0,1] simulation results of correlated seasonal innovations for wind, price and demand. To transform the uniform margins into realistic values, we use the quantiles of the trimmed residual series.

Using this series and the ARIMA model the simulation is conducted using the simulated innovations. The exponent of this result is combined with the new series to generate the plots (a)-(c), left from the red line is the original series and right from it the simulated series. The model clearly outperforms an ARIMA model with standard normal errors, that is not capturing any correlation between demand, wind and price, that the copula innovations are able to capture.

In table two the estimated ARIMA coefficients are shown, the best ARIMA Model is chosen according to its Aikaike estimation criteria. The model is assumed to be seasonal and we allow for models with non-zero mean. In order to achieve a positive simulation, we add the absolute minimum to the series, this does not change the character of the time series modeled but ensures consistent positive values.

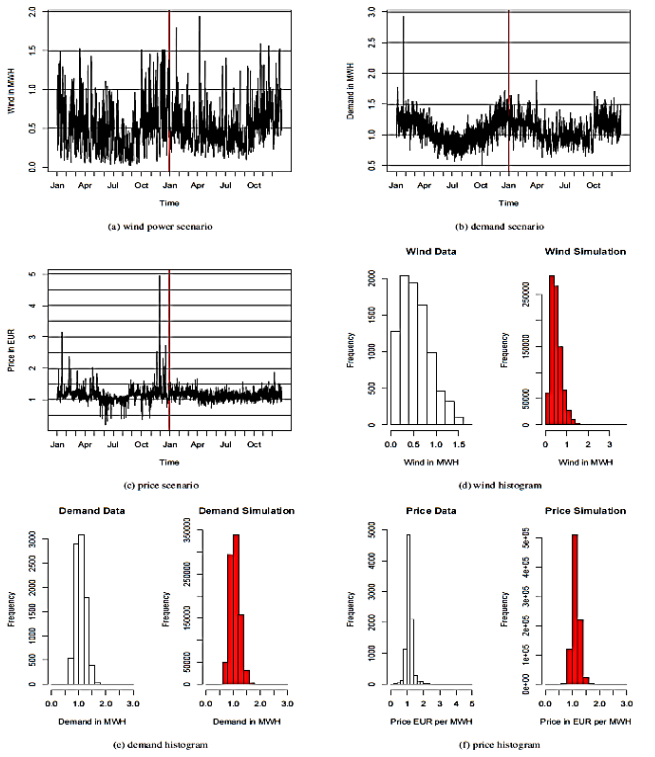


Fig. 1. Original data and generated scenarios for (a) wind power (MW/H) (b) demand (MW/H) (c) price (e) followed by subsequent distributions

## V. Discussions and Conclusion

Table one shows the kendall correlation of the empirical sample. Demand and price is positively correlated as well as demand and wind, we see a small negative relationship between wind and price, likely because the wind barely has influence on the price, outside of extremely windy circumstances. From the correlations themselves we cannot make conclusions about the endogeneity. Surprising is the large correlation between demand and wind and the lack thereof in terms of prices. But maybe when it is windy it is more likely cloudy, thus more energy is consumed for heat and light.

The proposed model with copulas can model dependencies, this benefit can be seen in table one, this table displays the range of the kendall correlation for all simulations. The range is wide, but it is reasonably close to the sample and is capturing a large portion of the observed sample correlation. The coefficients cannot be the same, because there is likely a higher correlation for extreme observations, which we omit for the simulation in order to receive more realistic simulations.

TABLE I: Data Correlation

| correlation sample | | | |
|---|---|---|---|
| | Demand | Price | Wind |
| Demand | 1 | | |
| Price | 0.4 | 1 | |
| Wind | 0.2 | -0.07 | 1 |
| correlation simulation | | | |
| Demand | 1 | | |
| Price | 0.14-0.31 | 1 | |
| Wind | 0.08-0.30 | 0-0.14 | 1 |

Table two shows the estimated ARIMA coefficients, since we required the model to be stationary, a mean is always estimated. This is reasonable here, because we attempt to forecast a year of data, because we just have a sample of one-year length, assuming there is a trend in the wind production would be likely overfitting the model in sample. The model is fitted on logarithms, to ensure positive values after the simulation. Below the values the standard errors are displayed.

TABLE II: Arima Coefficients

| ARIMA coefficient estimates Wind | | | | | |
|---|---|---|---|---|---|
| ar1 | ar2 | ma1 | | | mean |
| 1.5544 | -0.5705 | 0.1052 | | | 7.3758 |
| 0.0137 | 0.0136 | 0.0165 | | | 0.0558 |
| ARIMA coefficient estimates Demand | | | | | |
| ar1 | ar2 | ma1 | | | mean |
| 1.1107 | -0.1455 | -0.3716 | | | 6.9896 |
| 0.0681 | 0.0637 | 0.0652 | | | 0.0128 |
| ARIMA estimates Price | | | | | |
| ar1 | ar2 | ma1 | ma2 | ma3 | mean |
| 0.3117 | 0.493 | 1.1201 | 0.5782 | 0.2114 | 3.3671 |
| 0.0693 | 0.0624 | 0.0689 | 0.0381 | 0.0149 | 0.0073 |

The Fig. 1 shows each time series, on the left side of the vertical line is the original time series, on the right side the simulation. The time series is standardized to 1 and for the simulation we trim the values at 3%, this reduces the kurtosis of the residuals substantially and thus produces more reliable simulations over such a long-time frame. We tried different ranges and it produces still reasonable results. The histograms display that the sample properties are conserved,

we can see more outliers, because we have more observations in 100 simulations. The histograms show that the distribution of the year in sample and the simulations is reasonably close.

The model is able to capture correlation structures in the data that traditional approaches, like an ARIMA simulation with standard normal errors are not able to capture.

In this paper multivariate scenario generation based on three variables: demand, wind and price is presented. In the proposed multivariate scenario generation technique ARIMA is used for forecasting and copula for adjusting the residuals. The tail adjustment of the distribution and the impact is also discussed. In future works a comparative analysis of different statistical scenario generation technique for multivariate data would be conducted.

## REFERENCES


[1] P. Chen, T. Pedersen, B. Bak-Jensen, and Z. Chen, *IEEE Transactions on Power Systems*, vol. 25, no. 667, 2010.

[2] P. Schütz, A. Tomasgard, and S. Ahmed, *European Jour-nal of Operational Research*, vol. 199, no. 409, 2009.

[3] M. Kaut, *Computational Management Science*, vol. 11, no. 503, 2014.

[4] U. Schepsmeier, *Econometric Reviews*, pp. 1–22, 2016.

[5] G. Elidan, "Copula bayesian networks, in Advances in neural information processing systems," 2010, pp. 559–567

[6] T. Nagler, C. Schellhase, and C. Czado, *Dependence Modeling*, vol. 5, no. 99, 2017.

[7] C. Simard and B. Rémillard, *Dependence Modeling*, vol. 3, 2015.

[8] H. Hoeltgebaum, A. Street, and C. Fernandes, *IEEE Transactions on Power Systems*, 2018.

[9] M. Haugh, *Monte Carlo Simulation: IEOR EA703*, 2004.

[10] A. S. Czerny, *Astronomy and Astro-physics Supplement Series*, vol. 110, no. 405, 1995.

[11] D. Hakkarinen and Z. Chen, "Algorithmic cholesky fac-torization fault recovery, in parallel & distributed processing (IPDPS)," in *Proc. 2010 IEEE International Sym-posium on IEEE*, pp. 1–10, 2010.

[12] J. Dissmann, E. C. Brechmann, C. Czado, and D. Kurow-icka, *Computational Statistics and Data Analysis*, vol. 59, no. 52, 2013.



**Sambeet Mishra** received the bachelor's degree in electrical engineering and the master's degree with a specialization in power & energy systems. He has a PhD on Electrical Power Engineering.

His research interests are industrial mathematics (optimization), Data Science, and macroeconomics applications in power and energy systems.



**Christoph Würsig** graduated from the Leibniz University Hanover.

Currently he is working towards his doctorate degree. He is currently researching in the field of economics and finance with a focus on the commodity market.



**Chiara Bordin** received her master's degree in industrial engineering and her PhD in operational research at the University of Bologna (Italy). She worked as a researcher at the University of Durham (UK) and at the Norwegian University of Science and Technology (Norway).

She is currently a research scientist at SINTEF Energy Research, Norway. Her research interests are in the field of mathematical optimization applied to thermal and electrical energy systems.



**Ivo Palu** was born in Rakvere, Estonia, on July 26, 1979. He graduated from the Tallinn University of Technology (TalTech) and received the PhD degree in electrical power engineering in 2009. Currently he is a professor and director of department of electrical power engineering at TalTech. His subject of teaching is wind energy and electrical materials.

His field of research is wind turbine co-operation with thermal power plants and grid integration of new energy sources. He is a member of the board of Estonian Society for Electrical Power Engineering.