Long Range Capacity Planning

Ali Jalali
ajjalali@amazon.com
Amazon.com
Austin, Texas, USA

Pranesh Vyas
vyaspran@amazon.com
Amazon.com
Austin, Texas, USA

ABSTRACT

In this paper, we address the Long Range Capacity Planning Problem, which is to find percentile statistics of a demand time series over a distant future time window. Underestimating capacity impacts the quality of service to customers, and overestimating results in wasted resources and opportunities. Considering Auto Regressive Integrated Moving Average (ARIMA) as the model for time series forecasting, we show that standard training and inference consistently underestimates the true percentile statistics of the future demand time series, especially for long range forecasts. We propose modifying the training method and adding Monte Carlo simulation to the inference process. Our proposal reduces the percentile statistics forecast error to under 2% compared with 6% for the standard process, a 3x improvement. We present this result in the experiments section for both synthetic and real data.

CCS CONCEPTS

• **Applied computing** → **Forecasting**.

KEYWORDS

Auto Regressive Integrated Moving Average, ARIMA, Time Series Analysis, Forecasting, Capacity Planning, Long Range Forecasting, Quantile Regression, Percentile Analysis

ACM Reference Format:
Ali Jalali and Pranesh Vyas. 2022. Long Range Capacity Planning. In *Proceedings of KDD '22- Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '22).* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Recent events such as the pandemic, market volatilities and basic goods shortages have highlighted the role of machine learning, and in particular forecasting, in achieving operational excellence. It is vital for companies to be able to forecast and plan for future demand with a high degree of accuracy, sometimes months or years in advance. While the science community has made great strides forecasting averages, the problem of forecasting extreme scenarios has not been in the spotlight. In this paper, we address the issue of accurate extreme percentile forecasting. Specifically, we address Capacity Planning, a multi-faceted problem whose goal is to plan how much capacity (human resources, hardware, logistics, etc.) is required to meet a certain quality of service by taking a calculated risk in the relatively distant future, e.g., a few quarters or years ahead.

Unlike normal forecasting problems, Capacity Planning is not concerned with point-wise accuracy, but with population statistics such as percentiles across a (distant) period of time in the future. For example, assume an online retail company wants to plan their hardware requirements to be able to serve the surge of holiday season customers. Given that other competitors will also be in the market for more hardware, the company needs to plan ahead and acquire hardware early to ensure that it can meet its potential demand. At this point, the company faces a tradeoff between the wasted cost of unnecessary hardware and the risk of losing customers. One strategy is to proactively order hardware at e.g. the 95^{th} percentile of the potential customer distribution and leave the other 5%, if materialized, to reactive measures such as last minute hardware purchases. Capacity Planning is the proactive part of this strategy. It involves accurately forecasting the demand distribution over a period of time, estimating the percentile of this distribution that is consistent with tolerable risk, and planning for it.

In Capacity Planning, like many other forecasting problems, we need to model historical demand and forecast the future by making reasonable assumptions and modifications. Auto Regressive Integrated Moving Average (ARIMA) [4] is one of the classic time series forecasting models, and is proven to be optimal for weakly stationary time series. ARIMA can estimate a distribution for the future forecasts. The idea is to then get this distribution, find the percentile of interest and use it for planning.

Since its inception, many variations of ARIMA have been proposed to improve its forecast accuracy. Seasonal ARIMA (SARIMA), ARIMA with exogenous variables (ARIMAX), modeling innovation variance via Generalized Auto Regressive Conditional Heteroskedasticity (GARCH) [3] and Fuzzy ARIMA (FARIMA) [22] are just a few of the variations that have been proven to improve forecast accuracy under certain conditions. Another class of improvements to ARIMA is to create a hybrid model by cascading ARIMA with another model. Essentially, ARIMA provides a reliable hint for the second model, improving overall forecast accuracy. A few such models are ARIMA-ANN [29], ARIMA-QR [1], ARIMA-SVM [19] and ARIMA-Kalman [15]. All of these models use the same technique for parameter estimation, namely maximum likelihood with least squares.

A major limitation of ARIMA is that it can only forecast one step into the future, as the model depends on unobserved variables, which are unknown at future times. This makes long term forecasts of ARIMA, which are essential to solving the Capacity Planning problem, far less reliable. Furthermore, as we will discuss later,

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

long term ARIMA forecasts drift towards the mean of the time series. Since we are interested in one-sided percentiles for Capacity Planning, this results in underestimation of the demand percentiles for a given risk level.

Our contribution in this paper is to address these two shortcomings of ARIMA. We propose using Quantile Regression [21] for ARIMA parameter estimation instead of least squares, along with Monte-Carlo sampling during the forecasting process to avoid underestimation. We show that the combination of the two results in accurate estimation of the demand percentile for the given risk. Our experimental results on both synthetic and real data illustrate the accuracy of our proposed method against other alternatives.

Our results are based on our extensive experiments and while we discuss our intuition, we do not analyze our method theoretically. Furthermore, throughout the paper, we discuss the problem of capacity planning as a univariate time series forecasting problem for simplicity. The method can be generalized to multivariate time series forecasting case naturally.

The rest of the paper is organized as follows. In Section 3 we introduce our notation and definitions. Section 2 provides a formal definition of the Capacity Planning problem. We present our proposed solution in Section 4 and the experimental results in Section 5.

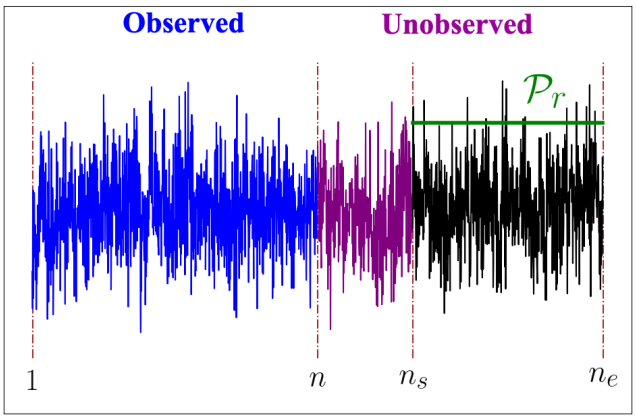

Figure 1: Capacity Planning Problem: Blue (between 1 and n) represents the observed part of the demand time series; Purple (between n and n_s) is the period between now and the start of the time window of interest; Black (between n_s and n_e) is the time window of interest for capacity planning. We are interested in forecasting $\mathcal{P}_r$ accurately to be able to plan the required capacity for the (one-sided) r^{th} percentile of the demand distribution. In other words, there chance of demand exceeding the planned capacity is $1-r$.

2 PROBLEM STATEMENT

In this section we formally define the capacity planning problem and its connection to the time series analysis. Capacity planning is a form of risk management, i.e., we want to guarantee a certain quality of a service with a certain risk level. For example, an online retail store wants to plan enough computer server capacity to be able to serve the surge of customers during holiday season. However, the exact number of customers is not known at the planning time and maximal planning for the absolute maximum number of customers, e.g. entire population of the country, not only is too expensive, but also is not realistic. One approach to this planning problem is to estimate some distribution for the number of customers and then use certain percentile of that distribution as the target for the planning. For instance, the online retail store manager might decide to take 5% risk; they then pick the 95^{th} percentile of the estimated distribution as their target demand and plan for it.

In order to estimate the distribution of the demand for a future time, it is very common to resort to historical trends of the demand and try to project those trends into the future. Furthermore, there are always unobserved factors that impact the demand both historically and in the future. The impact of these factors also need to be appropriately modeled. This process would help us get a distribution for a single point in the future; but this is not enough, because often, executing the planned capacity takes a long time. Imagine in the online retail store example, depending on the demand, we might need to build new data centers to accommodate the required computer servers. Moreover, due to overheads, we cannot repeat this process frequently, say build a new (small) data center every other month. Thus, a sustainable capacity planning should be able to plan early enough (to build the capacity) and for a wide enough window of time in the future (to minimize the overhead).

While overestimating the demand increases the quality of service, it also increases unnecessary costs on the system. In contrast, underestimation of the demand will result in customer dissatisfaction and ultimately impacts the bottom line. Thus, getting the percentile of demand that meets the tolerable risk is a critical task in capacity planning. Given the scale of the planning, even small deviations can translate to millions of dollars in cost or loss of customers. Below, we formally state the capacity planning problem.

Definition 1 (Capacity Planning Problem). *We are given a historical demand time series $x_1, x_2, \ldots, x_n$ where subscripts are time indices, denoted as $\mathbf{x}_1^n$, and, a tolerable risk level r for a period of n_s to n_e into the future. Suppose the time series takes values $\mathbf{x}_{n+1}^{n_e}$ in the future (which is unknown to us) and the empirical r^{th} percentile of the values of $\mathbf{x}_{n_s}^{n_e}$ (over the window of n_s to n_e) is $\mathcal{P}_r$. The capacity planning problem is to estimate the future demand percentile $\mathcal{P}_r(n_s, n_e)$ and plan for it. Figure 1 visualizes the capacity planning problem.*

A question might arise from this definition that if we plan our capacity for r^{th} percentile of the demand distribution, there is a $1-r$ chance that the demand exceeds the capacity and what happens in that case. Firstly, the choice of r involves cost benefit analysis of wasting resources versus degrading the quality of service for a small portion of demand and can be an executive choice. Secondly, the capacity planning is a proactive measure and there are always reactive measures built into systems to be able to handle an (relatively small) unexpected surge of demand for a limited time. These reactive measures can handle those $1-r$ percent of times. An example of a reactive measure for the online retail store could be a queuing mechanism that increases the response time to customers or an automatic scaling say within a cloud service.

Notice that for the capacity planning, we are not necessarily interested in accuracy of the prediction for each time index; rather, we are interested in the collective behavior of the time series within a time window of interest in the future. This is in contrast with typical objective of forecasting algorithms that try to minimize their point-wise error (on average). This is the key difference between the problem we are addressing here and normal forecasting problems that is studied in most other research works like [27].

3 DEFINITIONS & NOTATIONS

We start by introducing some notations and definitions. There are many ways to represent these concepts in the literature. We selected the representation that we think makes it easier to convey the message of this paper. We continue by defining Auto Regressive Moving Average (ARMA) as our main focus. We then explore other variations of ARMA and ultimately motivate why we limit the rest of the paper to ARMA and how the results can be expanded to those variations. We finally argue that the parameter estimation for these methods comes down to solving a least square problem. This is the key observation for our proposed solution.

3.1 ARMA & ARIMA

We represent a time series $x_1, x_2, \ldots, x_n$ where subscripts are time indices as $\mathbf{x}_1^n$. Considering practical applications, without loss of generality, we assume the observed part of time series starts from index 1 and ends at index n. Further, we assume that the time series is weakly stationary in the sense that with some differentiation process (as discussed below), the time series becomes stationary. In practice, this assumption means that the long term trend of the time series is a polynomial, e.g. there is a linear or quadratic (but not exponential) growth.

Definition 2 (ARMA(p, q)). *Given a* stationary *zero-mean time series $\mathbf{x}_1^n$, we model x_i for $i > n$ as*

$$\begin{aligned} x_i &= \mu_i + u_i \\ \mu_i &= \sum_{j=1}^{p} \phi_j x_{i-j} - \sum_{k=1}^{q} \theta_k u_{i-k} \\ u_i &\sim \mathcal{D}\left(0, \sigma^2\right) \end{aligned} \tag{1}$$

where, u_j's are i.i.d. samples drawn from the distribution $\mathcal{D}(0, \sigma^2)$, and, ϕ's, θ's and σ are the model parameters. Equivalently, we can represent ARMA as

$$\begin{aligned} x_i &\sim \mathcal{D}\left(\mu_i, \sigma_i^2\right) \\ \mu_i &= \sum_{j=1}^{p} \phi_j \mu_{i-j} - \sum_{j=1}^{p} \phi_j u_{i-j} - \sum_{k=1}^{q} \theta_k u_{i-k} \\ \sigma_i^2 &= \sigma^2 \end{aligned} \tag{2}$$

This representation formulates the evolution of the mean of the distribution while keeping the variance constant. It models the time series as samples drawn from the distribution $\mathcal{D}$ with the mean of μ_i and constant standard deviation of σ.

Given a stationary time series $\mathbf{x}_1^n$, we use the auto-ARIMA process [12] to find the optimal parameters p and q. This process is independent of the methodology we are using for estimating parameters ϕ, θ and σ. Throughout the paper, we assume that p and q are given and we focus on different methodologies for parameter estimation.

Definition 3 (ARIMA(p, d, q)). *Given a (non-stationary) time series $\mathbf{x}_1^n$, suppose we apply the differentiation process d times to get a stationary time series $\mathbf{y}_1^{n-d}$. We then define*

$$ARIMA(p, d, q)\left(\mathbf{x}_1^n\right) = ARMA(p, q)\left(\mathbf{y}_1^{n-d}\right) \tag{3}$$

where, ARMA is defined in Definition 2.

Based on this definition, by doing a pre-processing on the time series, we can reduce an ARIMA model to an ARMA model. In practice, there are many stationarity tests such as KPSS [14], ADF [10], PP [20], and, ADF-GLS [11], that can be used to determine the value of d. Once d is determined, one can run the differentiation process and reduce the problem to parameter estimation for the ARMA model. For other variations of the ARMA, please see the appendix.

3.2 Model Parameter Estimation

There are many methods in the literature for estimating the parameters of the models introduced in this section. We categorize these methods into few groups and discuss the capabilities and limitations of each group. While we only discuss the core idea of this paper in the context of two-step regression (due to simplicity) and maximum likelihood estimation (due to optimality), the idea can be generalized to other methods. Furthermore, we only consider the parameter estimation for ARMA(p, q) model with distribution $\mathcal{D}$ being normal distribution for the ease of notation. Again, the methods can be easily generalized to other models.

The main difficulty of parameter estimations for these models arises from the fact that the innovations u_i's are not observed. Different methods need to either explicitly or implicitly, iteratively or statically estimate these unobserved variables. Once we have an estimate for the unobserved variables, it is not hard to see that all of the models introduced above are some form of linear regression (or bilinear regression [24]). The core idea of this paper is to perform this linear regression using different criteria to address the capacity planning problem. We will formally define the problem in Section 2.

3.2.1 Two-Step Regression. In this method, we first assume the innovations u_i's are all zero and do a regression on $\mathbf{x}_1^n$ by using the Definition 1 to get an ordinary least squares estimate for $\hat{\phi}$'s from the system of $n - p$ equations

$$x_i = \sum_{j=1}^{p} \hat{\phi}_j x_{i-j} \qquad \forall i \in \{p+1, p+2, \ldots, n\} \tag{4}$$

We then can estimate the unobserved innovations $u_i = x_i - \sum_{j=1}^{p} \hat{\phi}_j x_{i-j}$. Finally, using x_i's and u_i's, we can do a second ordinary least squares to estimate ϕ's and θ's from the system of $n - \max(p, q)$ equations

$$x_i = \sum_{j=1}^{p} \phi_j x_{i-j} - \sum_{k=1}^{q} \theta_j u_{i-j} \qquad \forall i \in \{\max(p, q) + 1, \ldots, n\} \tag{5}$$

Subsequently, the parameter σ can be estimated by

$$\widehat{\sigma^2} = \frac{1}{n - \max(p,q) - 1} \sum_{i=\max(p,q)+1}^{n} u_i^2 \tag{6}$$

This method is very fast in terms of run time and easy to understand, but unfortunately it is not robust and the accuracy might not be great under certain conditions [16].

3.2.2 Method of Moments. The idea of the method of moments is to find a distribution that matches the moments of the data. Dominantly the Yule-Walker algorithm [23, 25] is used to estimate the moments of the time series. The parameters of the ARMA model can then be estimated by solving non-linear system of equations that ties them to the moments. This method is computationally intense and it might not converge under certain conditions to the extent that some researchers suggested this method should not be used for ARMA parameter estimation [7].

3.2.3 Maximum Likelihood. Maximum likelihood maximized the likelihood function for the distribution $\mathcal{D}$ to find the optimal parameters of the ARMA model. Conditioned the first $\max(p,q)$ elements of the time series, i.e., $\mathbf{x}_1^{\max(p,q)}$, we can write the joint probability density function of $\mathbf{x}_1^n$ as

$$\begin{aligned} &f\left(\mathbf{x}_1^n; \phi, \theta, \sigma\right) = \\ &\quad f\left(\mathbf{x}_1^{\max(p,q)}; \phi, \theta, \sigma\right) f\left(\mathbf{x}_{\max(p,q)+1}^n | \mathbf{x}_1^{\max(p,q)}; \phi, \theta, \sigma\right) \end{aligned} \tag{7}$$

where, $f(.)$ represents the probability density function. Hence, the log-likelihood function can be written as

$$\begin{aligned} \mathcal{L}\left(\mathbf{x}_1^n; \phi, \theta, \sigma\right) = &\overbrace{-\log\left(f\left(\mathbf{x}_1^{\max(p,q)}; \phi, \theta, \sigma\right)\right)}^{\triangleq \mathcal{L}_{\mathbf{m}}\left(\mathbf{x}_1^{\max(p,q)}; \phi, \theta, \sigma\right)} \\ &\underbrace{-\log\left(f\left(\mathbf{x}_{\max(p,q)+1}^n | \mathbf{x}_1^{\max(p,q)}; \phi, \theta, \sigma\right)\right)}_{\triangleq \mathcal{L}_{\mathbf{c}}(\mathbf{x}_1^n; \phi, \theta, \sigma)} \end{aligned} \tag{8}$$

where, $\mathcal{L}(.)$ represents the negative log likelihood function, and, $\mathcal{L}_{\mathbf{c}}(.)$ represents the conditional negative log likelihood function, and, $\mathcal{L}_{\mathbf{m}}(.)$ represents the marginal negative log likelihood function. Assuming $\mathcal{D}$ to be the normal distribution, the conditional maximum likelihood estimation of ARMA model comes down to minimizing the conditional least squares, i.e.,

$$\begin{aligned} \min\ \mathcal{L}_{\mathbf{c}}\left(\mathbf{x}_1^n; \phi, \theta, \sigma\right) \ &\sim\ \min\ \mathrm{S}_{\mathbf{c}}\left(\mathbf{x}_1^n; \phi, \theta, \sigma\right) \\ &= \min_{\phi,\theta} \sum_{i=\max(p,q)}^{n} \left(x_i - \sum_{j=1}^{p} \phi_j x_{i-j} - \sum_{k=1}^{q} \theta_j u_{i-j}\right)^2 \end{aligned} \tag{9}$$

and σ can be estimated via (6). In order to solve (9), one can use the iterative method of Algorithm. 1.

Characterizing $\mathcal{L}_{\mathbf{m}}(.)$ is done via writing each x_i for $1 \le i \le \max(p,q)$ as an infinite sum of innovations u_j and find the covariance matrix of $x_1, \ldots, x_{\max(p,q)}$. We skip presenting all the details here and refer to [9], because it has been shown that $\mathcal{L}(.)$ and $\mathcal{L}_{\mathbf{c}}(.)$ have the same asymptotic behavior. In other words, for sufficiently large values of n, given the simplicity and lower computational cost, one can optimize $\mathcal{L}_{\mathbf{c}}(.)$ and skip the full likelihood function.

This method is consistent, asymptotically normally distributed and asymptotically efficient. Also, it can also be used to detect and remove outliers [5, 6, 17]. Most popular implementations of ARMA parameter estimation used in the industry are based on the maximum likelihood method with a normal distribution assumption.

The two-step regression and conditional maximum likelihood methods are least square optimizations at their core. These methods can be extended to seasonal and GARCH versions of ARIMA as they have a similar structure (see appendix). For the case of seasonality, the least square problem becomes a bilinear problem that can be solved by alternating between the parameters as described earlier.

4 MAIN CONTRIBUTION

In this section we discuss our proposal for the Capacity Planning Problem defined in Definition 1. We start by reviewing the existing solutions and their shortcomings and then move onto our contribution. In short, our proposal involves changing the objective function for ARIMA parameter estimation from the least squares optimization to the Quantile Regression (QR) [13], and, using Monte-Carlo simulation to model innovations. We discuss why this is the right choice for the Capacity Planning Problem and show the experimental results in Section 5.

In general, ARIMA model (along with its variations discussed in Section 3) is a good candidate to model the historical and future demands for the Capacity Planning Problem, because it accounts for unobserved factors in the modeling and (if necessary) is capable of handling the evolution of mean and variance of these latent factors over time. In fact, the optimality of ARMA for stationary time series is established by Wold's Decomposition Theorem [28] and as long as the non-stationarity of the time series is polynomial, the differentiation process of ARIMA can convert it to a stationary process. This justifies our focus on ARIMA model among other proposals in the literature.

4.1 Existing Solutions

ARIMA model forecasts the standard deviation $\hat{\sigma}$ of innovations u_i. Assuming the distribution of innovations is a normal, one can use the error function to compute the desired percentile. For example, if we are interested in $r = 97.5\%$, we have $\mathcal{P}_r \approx \hat{x}_i + 2\hat{\sigma}$ at any point i.

The main issue with this method as illustrated in Fig. 2 is that while ARMA is optimal to forecast a single point into the future, its long term forecast drifts towards the mean of the time series (that is modeled by the intercept in the model). This happens mainly because there is no way to estimate the unobserved innovations u_i during forecasting other than replacing them with the best blind estimate which is the mean of the distribution $\mathcal{D}$, i.e., zero. Hence, after q step forecast the Moving Average (MA) part of the model becomes zero. Subsequently, since the Auto Regressive (AR) part is stable, it drifts towards the zero and the forecast becomes the constant in the mode, i.e., the intercept. Now, since $\mathcal{P}_r$ is just a constant away from this forecast and we are interested in one-sided percentiles for the Capacity Planning Problem, this phenomenon results in under-estimation of the target parameter $\mathcal{P}_r$.

In the case of ARIMA (or other variations of ARMA), depending on the value of d, the forecast becomes a polynomial. For example,

Algorithm 1: Solve Conditional Maximum Likelihood

ConditionalMLE *(x, p, q)*

Solve system of equations $x_i = \sum_{j=1}^{p} \hat{\phi}_j x_{i-j}$ for all $i \in \{p+1, \ldots, 2p\}$ to get $\hat{\phi}$'s

Initialize $u_i = 0$ for all $i \in \{1, \ldots, p\}$ and $u_i = x_i - \sum_{j=1}^{p} \hat{\phi}_j x_{i-j}$ for all $i \in \{p+1, \ldots, n\}$

while *Not Converged* **do**

  Regress via least squares: $x_i = \sum_{j=1}^{p} \phi_j x_{i-j} - \sum_{k=1}^{q} \theta_k x_{i-k}$ for all $i \in \{\max(p,q)+1, \ldots, n\}$

  Update $u_i = x_i - \sum_{j=1}^{p} \phi_j x_{i-j} - \sum_{k=1}^{q} \theta_k x_{i-k}$ for all $i \in \{1, \ldots, n\}$

return ϕ, θ

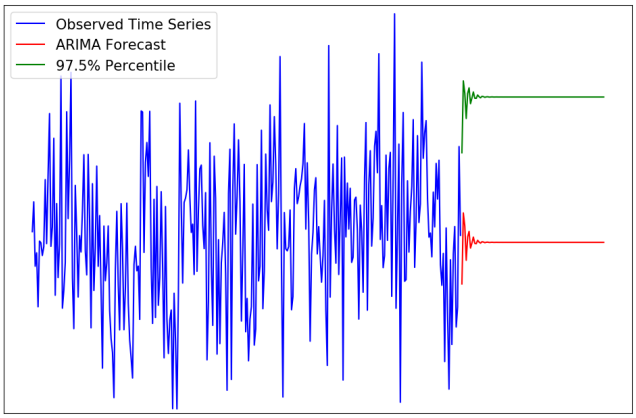

Figure 2: Long-term forecasting with ARMA model drifts towards the mean of the time series, i.e., the intercept used in the model. Thus, the percentiles of the presumed normal distribution also becomes flat. Since Capacity Planning Problem is only interested in one-sided percentiles, this results in under-estimating the true percentiles.

if $d = 1$, the forecast becomes a straight line with a constant slope. The slope is equal to the intercept in the underlying ARMA model. Again, due to the fact that we are interested in one-sided percentiles, this method will under-estimate $\mathcal{P}_r$ in long term forecasting.

4.2 Proposal

Our proposal has two ingredients based on the specific requirements of the Capacity Planning Problem. One ingredient is replacing the least squares in the ARMA parameter estimation with quantile regression, and, the other ingredient is the simulation of innovations during the forecasting process. We discuss these two ingredients first and then uncover our proposal which is the mixture of the aforementioned two ingredients.

4.2.1 Quantile Regression. The least squares method for ARMA parameter estimation as discussed in Section 3.2 aims to minimize the average of the square of errors. If we consider the mean of the time series as a random variable, then the least squares converges to the 50^{th} percentile of that random variable. This value, i.e., the intercept, is what the ARMA forecast drifts towards as discussed in Section 4.1. Now, when we add e.g. $2\hat{\sigma}$ to this intercept to get to the 97.5^{th} percentile, in half of the realizations, this falls short of the true 97.5^{th} percentile. If instead of 50^{th} percentile for the intercept, we had a way to estimate 97.5^{th} percentile of the intercept and then add e.g. $2\hat{\sigma}$ to this new mean, that would put us closer to the true 97.5^{th} percentile.

Quantile regression is a method that allows us to target a desired percentile of a random variable (as opposed to 50^{th} percentile with least squares). The objective of the quantile regression is

$$\min_{\phi,\theta} \sum_{i=\max(p,q)}^{n} \rho_r \left(x_i - \sum_{j=1}^{p} \phi_j x_{i-j} + \sum_{k=1}^{q} \theta_k x_{i-k} \right) \tag{10}$$

where, $\rho_r(.)$ is defined as $\rho_r(z) = (1-r)\max(z, 0) + r\max(-z, 0)$. This minimization converges to the point where r percentile of the error mass falls below and $1-r$ percentile of the error falls above it. Thus, the first ingredient of our proposal is the use of Quantile Regression instead of least squares.

4.2.2 Monte-Carlo Simulations. During forecasting procedure using ARIMA model, we forecast the first next step using the historical data, but after that, we assume the unobserved innovations u_i's are all zero and practically remove the MA part of the ARMA. Since Capacity Planning Problem is impacted by the long term population statistics and is not directly impacted by the point accuracy of the forecasts, we can simulate innovations based on the estimated $\hat{\sigma}$.

The modified forecasting process works as follows: we first forecast one step into the future; we then get a sample from the normal distribution with estimated $\hat{\sigma}$ and add it to the forecast pretending we had the full knowledge of the innovation u_i. We then continue the forecasting process, one step at a time, as explained to reach n_e. One can repeat this whole process multiple times to get a better estimate of $\mathcal{P}_r$ over the simulated time series. We will present the results of such simulations in Section 5.

4.2.3 Capacity Planning Proposal. Putting the two ingredients together, we propose changing ARMA parameter estimation by replacing least squares with quantile regression, and, doing Monte-Carlo simulation for innovation ensembles, say K times, to get an accurate estimate of $\mathcal{P}_r$. In particular, in our experience, the average of 3rd and 1st quantiles of the simulated percentiles is the best candidates as we will see in Section 5. Algorithm 2 summarizes our proposal.

The choice of quantile regression with Monte-Carlo simulation introduces large variance in our forecast due to the fact that quantile regression is very sensitive to the single data point that represents the r^{th} quantile. That is why the choice of the average of first and third quartiles for our final estimate is crucial. Essentially, this averaging assumes that between first and third quartiles, our

Algorithm 2: Solve Capacity Planning Problem

CapPlan *(x, p, q, r, K)*

Replace least squares with quantile regression in Algorithm 1
for *i* **from** 1 **to** *K* **do**
  Forecast ARIMA time series for $n_e - n$ steps
  Compute and store the $\mathcal{P}_r$ of x_i for $i \in \{n_s, \ldots, n_e\}$
return *Average of 3rd and 1st quantile of* $\mathcal{P}_r$*'s*

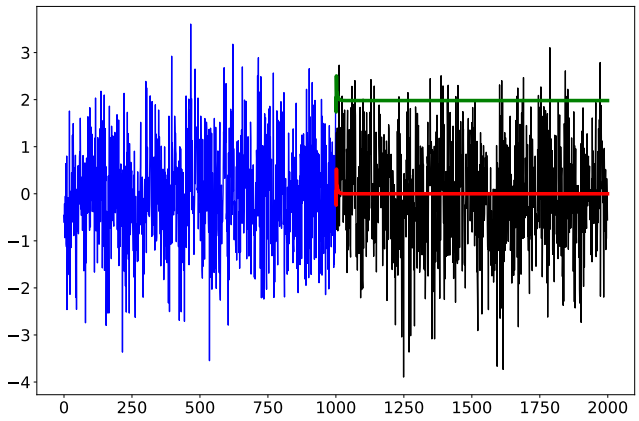

Figure 3: Usual ARMA Forecast: Blue represents the observed Time-series (Train), Black represents the unobserved Time-series (Test), Red is the ARMA Forecast and Green is the 97.5^{th} percentile obtained by adding $2\hat{\sigma}$ to the forecast.

parameter is symmetrically distributed and hence, the average is a better estimate of the true median (than the observed median).

5 EXPERIMENTAL RESULTS

In this section, we examine the accuracy of our proposed method for the estimation of $\mathcal{P}_r$ in comparison to the alternatives. We consider four different methods over two dimensions: ARIMA with least squares vs ARIMA with quantile regression; and; using Monte-Carlo sampling vs not. We also present the result for two-step regression (see Section 3.2.1) and conditional maximum likelihood (see Section 3.2.3). Results are presented for both synthetic and real data for each case.

For the sake of the consistency across the experiments, we fix $r = 97.5\%$, i.e., we examine how close $\mathcal{P}_r$ is to the 97.5^{th} percentile of the data. This choice makes it relatively easy to compute $\mathcal{P}_r$ for a normal distribution because it is just 2σ above the mean of the distribution. Again, notice that we are not concerned about the point to point fluctuations for Capacity Planning purposes as long as the premise is satisfied.

5.1 Synthetic Data

We generate 1000 independent sets of synthetic data of length 2000 using ARMA(3, 2) with parameters $\phi = [0.3, -0.2, 0.4]$, $\theta = [-0.2, 0.1]$ and $\sigma = 1.0$. For each set, we use the first 1000 data points for training and the second 1000 data points for testing. For

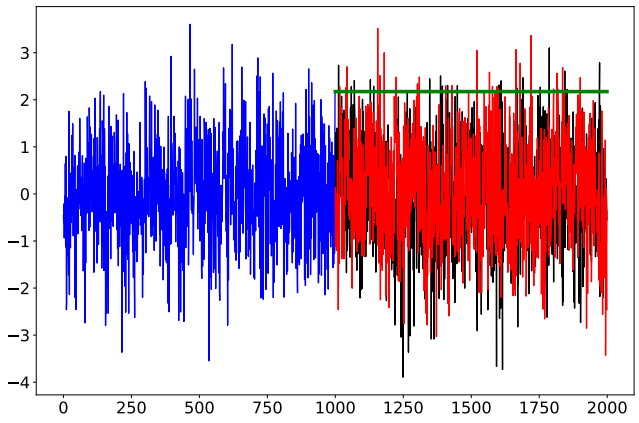

Figure 4: Monte-Carlo Sampling of Innovations Forecast: Blue represents the observed Time-series (Train), Black represents the unobserved Time-series (Test), Red is the ARMA forecast whose innovations are simulated by sampling from the normal distribution $\mathcal{N}(0, \hat{\sigma})$ and Green is the 97.5^{th} percentile of the Red (that needs to be compared against the 97.5^{th} percentile of the Black for accuracy).

each set of synthetic data and each modeling method, we estimate $\widehat{\mathcal{P}_r}$ and record the count of data points from the test data that fall under this value. Since the length of the test data is 1000 and we aim at 97.5^{th} percentile, ideally we expect 975 test data fall under $\widehat{\mathcal{P}_r}$ and 25 fall above. Thus, we measure ourselves against the ideal number of 975.

For each set of data, we do four different parameter estimation methods (via two-step regression or conditional maximum likelihood; and; via least squares and quantile regression), and, two different inference/forecasting methods (with Monte-Carlo simulation of innovations and without). Figure 3 illustrates an inference/forecasting without Monte-Carlo simulation. In this case $\mathcal{P}_r$ is computed by adding $2\hat{\sigma}$ to the model forecast. In contrast, Figure 4 depicts an inference/forecasting using Monte-Carlo simulation of the innovations. In this case, after forecasting, we compute the sample 97.5^{th} percentile and count that as $\mathcal{P}_r$.

Figures 5 & 6 show the box plot of the number of test data points falling under the estimated $\mathcal{P}_r$ for two-step regression and conditional maximum likelihood respectively against the ideal number 975 that is shown with a green line. In both cases, forecasting without sampling tend to under-estimate $\mathcal{P}_r$ since the green line is above the 3rd quartile. Furthermore, in both cases, using standard ARMA with sampling, over-estimates the $\mathcal{P}_r$ and only our proposal, which is the mix of quantile regression and Monte-Carlo sampling has the optimal number within its 1st and 3rd quartiles. Practically, we found that the average of the 1st and 3rd quartiles is the best estimator of $\mathcal{P}_r$ and in particular, median over-estimates $\mathcal{P}_r$. Furthermore, it seems that the results of conditional maximum likelihood are just 10% closer to the ideal point 975. Given the higher computational complexity of the conditional maximum likelihood versus the two-step regression, one might decide to choose one over the other based on the trade-offs.

Dataset	MLE-LR	MLE-LR-Sampling	MLE-QR	MLE-QR-Sampling	$\mathcal{P}_r$
NY COVID-19 [2]	15573	17006	15632	**16421**	16648
CO Air Quality [8]	1529	1693	1570	**1655**	1622
US CO2 Emission [18]	5,780,681	5,950,062	5,812,423	**5,902,848**	5,879,655
Daily Delhi Climate [26]	36.41	37.85	36.62	**37.15**	37.25

Table 1: Comparison of the estimated $\widehat{\mathcal{P}_r}$ using different methods against the actual $\mathcal{P}_r$. Here we used conditional maximum likelihood since it gives better results for all methods. This result shows that our proposal under QR-Sampling column is superior to other methods in forecasting the parameter $\mathcal{P}_r$ for Capacity Planning.

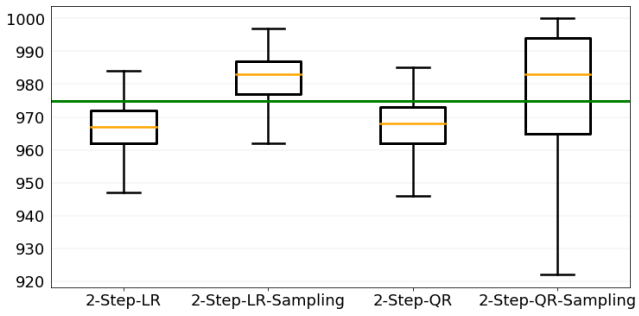

Figure 5: Synthetic Data - Distribution of the number of test data falling below the estimated $\widehat{\mathcal{P}_r}$ for two-step regression with least squares and quantile regression and with/without sampling over 1000 runs. The method with quantile regression and sampling gives the best performance as the ideal number 975 (depicted with a green line) falls between its 1st and 3rd quarter. The average of 1st and 3rd quantiles seem to be the best estimate for $\widehat{\mathcal{P}_r}$.

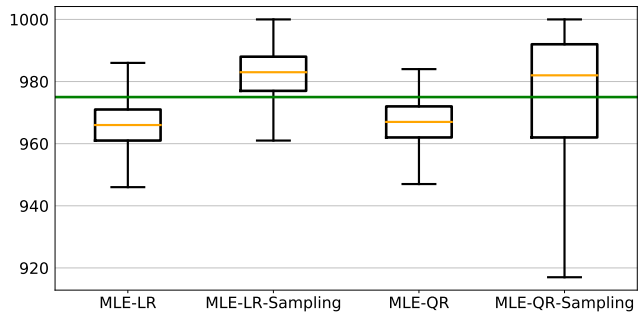

Figure 6: Synthetic Data - Distribution of the number of test data falling below the estimated $\widehat{\mathcal{P}_r}$ for conditional maximum likelihood with least squares and quantile regression and with/without sampling over 1000 runs. The method with quantile regression and sampling gives the best performance as the ideal number 975 (depicted with a green line) falls between its 1st and 3rd quarter. The average of 1st and 3rd quantiles seem to be the best estimate for $\widehat{\mathcal{P}_r}$.

It appears that our proposed method results in higher variance in the outcome. The question arises if this method is robust when we consider the average of first and third quartiles. We repeated the entire process, including the generation of data using the same parameters, 1000 times and measured the variance on the final estimate. We observed less than 2% variance (1.83% to be exact) in estimating $\mathcal{P}_r$. This shows that while quantile regression sensitivity increases the variance in Monte-Carlo simulation outcomes, it does not have a big impact on our final estimate.

5.2 Real Data

We examine our proposed method on some public data sets and report the results. We split each data set into 75% train and 25% test and use only the train portion for ARIMA parameter estimation. Similar to the synthetic data experiments, we set $r = 97.5\%$. The estimated $\widehat{\mathcal{P}_r}$ for the cases that we do not simulate the innovations is just a single number (steady-state mean plus $2\hat{\sigma}$); while it is a distribution when we repeat the forecasting process 1000 times by Monte-Carlo sampling of the innovations. In the latter case, we use the average of 1st and 3rd quantiles as our final estimate for $\widehat{\mathcal{P}_r}$. We also compute the actual $\mathcal{P}_r$ from the test data and report it for comparison.

We used four real data sets for this experiment. The NY COVID-19 [2] data set is the number of daily positive COVID-19 tests in the state of NY from March 2020 till Feb 2021. The underlying data is broken down by county and we aggregated them into the state level. The CO Air Quality [8] is a UCI data set collected in a city in Italy. The data set has hour-by-hour level of carbon monoxide in the air. We created a time series by concatenating all of the data. The US CO2 Emission [18] data set is a relatively small data set that recorded annual CO2 emission for all countries since 1970 till 2016. We used the CO2 emission of the United States and created a time series. Finally, Daily Delhi Climate [26] has recorded the daily temperature of Delhi in Celsius from 2013 to 2017.

Table 1 summarizes the estimated $\widehat{\mathcal{P}_r}$ for different data sets against the actual $\mathcal{P}_r$ measured from the test data. In all cases, our proposed method is better (or similar) to other alternatives. The results shows that the error in estimation of $\mathcal{P}_r$ with our proposed method is under 2% while the error for standard ARIMA is over 6%, i.e. $3x$ worse.

6 CONCLUSION

We presented a formal definition for the Capacity Planning Problem which is a crucial planning step for many companies given the volatilities in the marketplace. We showed how existing methods

for forecasting percentiles fail to produce an accurate estimate especially when we do long range forecasting. We then proposed a two-ingredient solution to the problem by modifying the parameter estimation process for ARIMA and adding a Monte-Carlo simulation sampling to the inference/forecasting process. Although we discussed in the context of ARMA for most cases, we explained how this proposal can be expanded to other variations like ARIMA, SARIMA, GARCH, etc. With our experiments, we demonstrated that the proposed solution is effective and makes the estimation of percentiles very accurate even for long range forecasts.

Future works can include theoretical analysis of our proposed solution and expanding it to other areas of density estimation. Another interesting direction is to analyze whether using old data is hurting our forecasts due to the fact that the evolution of time series might have completely altered over time and a single model might not describe the entire history of time series.

ACKNOWLEDGMENTS

To Amazon Prime Video Forecasting and Capacity Planning team for the support, discussions and comments that made this paper possible.

REFERENCES

[1] Wiwik Anggraeni, Faizal Mahananto, Fajar Ratna Handayani, A. Kuntoro Boga, and Sumaryantoe. 2019. Hybrid of ARIMA and Quantile Regression (ARIMA-QR) Model for Forecasting Paddy Price in Indonesia. *Journal of Engineering and Applied Sciences* 14 (2019), 7609–7619.
[2] Domenico Benvenuto, Marta Giovanetti, Lazzaro Vassallo, Silvia Angeletti, and Massimo Ciccozzi. 2020. Application of the ARIMA model on the COVID-2019 epidemic dataset. *Data in brief* 29 (2020), 105340.
[3] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31, 3 (1986), 307–327.
[4] George EP Box, Gwilym M Jenkins, Gregory C Reinsel, and Greta M Ljung. 2015. *Time series analysis: forecasting and control.* John Wiley & Sons.
[5] Oscar H Bustos and Victor J Yohai. 1986. Robust estimates for ARMA models. *J. Amer. Statist. Assoc.* 81, 393 (1986), 155–168.
[6] Yacine Chakhchoukh. 2010. A new robust estimation method for ARMA models. *IEEE Transactions on Signal Processing* 58, 7 (2010), 3512–3522.
[7] MJL De Hoon, THJJ Van der Hagen, H Schoonewelle, and H Van Dam. 1996. Why Yule-Walker should not be used for autoregressive modelling. *Annals of nuclear energy* 23, 15 (1996), 1219–1228.
[8] Saverio De Vito, Ettore Massera, Marco Piga, Luca Martinotto, and Girolamo Di Francia. 2008. On field calibration of an electronic nose for benzene estimation in an urban pollution monitoring scenario. *Sensors and Actuators B: Chemical* 129, 2 (2008), 750–757.
[9] Warren Dent. 1977. Computation of the exact likelihood function of an ARIMA process. *Journal of Statistical Computation and Simulation* 5, 3 (1977), 193–206.
[10] David A Dickey and Wayne A Fuller. 1979. Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American statistical association* 74, 366a (1979), 427–431.
[11] Graham Elliott, Thomas J Rothenberg, and James H Stock. 1992. *Efficient tests for an autoregressive unit root.* Technical Report. National Bureau of Economic Research.
[12] Rob J Hyndman and Yeasmin Khandakar. 2007. *Automatic time series for forecasting: the forecast package for R.* Clayton VIC, Australia: Monash University, Department of Econometrics and Business Statistics.
[13] Roger Koenker and Kevin F Hallock. 2001. Quantile regression. *Journal of economic perspectives* 15, 4 (2001), 143–156.
[14] Denis Kwiatkowski, Peter CB Phillips, Peter Schmidt, and Yongcheol Shin. 1992. Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of econometrics* 54, 1-3 (1992), 159–178.
[15] Hui Liu, Hong-qi Tian, and Yan-fei Li. 2012. Comparison of two new ARIMA-ANN and ARIMA-Kalman hybrid methods for wind speed prediction. *Applied Energy* 98 (2012), 415–424.
[16] Colin R Mckenzie. 1997. The properties of some two step estimators of ARMA Models. *Mathematics and computers in simulation* 43, 3-6 (1997), 451–456.
[17] Nora Muler, Daniel Pena, Victor J Yohai, et al. 2009. Robust estimation for ARMA models. *The Annals of Statistics* 37, 2 (2009), 816–840.
[18] Jos GJ Olivier, Greet Janssens-Maenhout, Marilena Muntean, and Jeroen AHW Peters. 2016. Trends in global CO2 emissions: 2016 Report. European Commission. *Joint Research Centre (JRC), Directorate C–Energy, Transport and Climate* (2016).
[19] Ping-Feng Pai and Chih-Sheng Lin. 2005. A hybrid ARIMA and support vector machines model in stock price forecasting. *Omega* 33, 6 (2005), 497–505.
[20] Peter CB Phillips and Pierre Perron. 1988. Testing for a unit root in time series regression. *Biometrika* 75, 2 (1988), 335–346.
[21] James W Taylor and Derek W Bunn. 1999. A quantile regression approach to generating prediction intervals. *Management Science* 45, 2 (1999), 225–237.
[22] Fang-Mei Tseng and Gwo-Hshiung Tzeng. 2002. A fuzzy seasonal ARIMA model for forecasting. *Fuzzy Sets and Systems* 126, 3 (2002), 367–376.
[23] George Udny Yule. 1927. On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London Series A* 226 (1927), 267–298.
[24] Dietrich Von Rosen. 2018. Bilinear regression analysis. *Lecture notes in statistics* 220 (2018).
[25] Gilbert Thomas Walker. 1931. On periodicity in series of related terms. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 131, 818 (1931), 518–532.
[26] Team WeatherUnderground. 2018. Daily climate data in the city of Delhi from 2013 to 2017. https://www.kaggle.com/sumanthvrao/daily-climate-time-series-data
[27] Ruofeng Wen, Kari Torkkola, Balakrishnan Narayanaswamy, and Dhruv Madeka. 2017. A multi-horizon quantile recurrent forecaster. *arXiv preprint arXiv:1711.11053* (2017).
[28] Herman Wold. 1938. *A study in the analysis of stationary time series.* Ph.D. Dissertation. Almqvist & Wiksell.
[29] G Peter Zhang. 2003. Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing* 50 (2003), 159–175.

A SEASONALITY IN ARIMA

We define seasonality with periodicity of m to be a consistent pattern that can be seen in a given time series $\mathbf{x}_1^n$ if we consider any sub-series of the form $x_i, x_{i+m}, x_{i+2m}, \ldots$ where $i \in \{1, 2, \ldots, m-1\}$. For example, if x_i represents daily numbers and $m = 7$, i.e., we have weekly patterns, then $x_1, x_8, x_15, \ldots$ represents the sequence of Sundays and $x_2, x_9, x_16, \ldots$ represents the sequence of Mondays, etc.

Definition 4 (SARMA$(p,q)(P,Q)(m)$). *Given a stationary zero-mean time series $\mathbf{x}_1^n$ with seasonal periodicity of m, we model x_i for $i > n$ as*

$$\begin{aligned} x_i &= \mu_i + u_i \\ \mu_i &= \sum_{j=1}^{p} \phi_j x_{i-j} + \sum_{l=1}^{P} \Phi_l x_{i-lm} - \sum_{j=1}^{p}\sum_{l=1}^{P} \phi_j \Phi_l x_{i-j-lm} \\ &\quad - \sum_{k=1}^{q} \theta_k u_{i-k} - \sum_{v=1}^{Q} \Theta_v u_{i-vm} + \sum_{k=1}^{q}\sum_{v=1}^{Q} \theta_k \Theta_v u_{i-k-vm} \\ u_i &\sim \mathcal{D}\left(0, \sigma^2\right) \end{aligned} \tag{11}$$

where, u_j's are i.i.d. samples drawn from the distribution $\mathcal{D}(0, \sigma^2)$, and, ϕ's, Φ's, θ's, Θ's and σ are the model parameters.

In contrast to ARMA (1), SARMA has extra terms to capture the seasonality. The presence of these terms breaks the linearity of μ_i in model parameters. However, later we leverage the fact that given θ's and ϕ's, the μ_i is linear in Θ's and Φ's and vice versa.

Definition 5 (SARIMA$(p,d,q)(P,D,Q)(m)$). *Given a (non stationary) time series $\mathbf{x}_1^n$ with seasonal periodicity of m, suppose we apply the seasonal differentiation process D times (via Algorithm 3) followed by the application of the differentiation process d times to*

Algorithm 3: Seasonal Differentiation Process

SeasonalDiff (**x**, D, m)

if $d \leq 0$ **then**
  return x
for i **from** 1 **to** **x**.$length - m$ **do**
  $y_i = x_{i+m} - x_i$
return SeasonalDiff (**y**, $D - m$, m)

get a stationary time series $\mathbf{y}_1^{n-d-mD}$. *We then define*

$$SARIMA(p,d,q)(P,D,Q)(m)\left(\mathbf{x}_1^n\right) = SARMA(p,q)(P,D,Q)(m)\left(\mathbf{y}_1^{n-d-mD}\right) \tag{12}$$

where, SARMA is defined in Definition 4.

Similar to the ARIMA case, we can focus on SARMA instead of SARIMA given that the appropriate pre-processing (differentiation) has taken place. Furthermore, we assume the parameters p, q, P and Q are known perhaps via an auto-ARIMA process as explained in the previous section. The choice of the parameter m can be done either via domain knowledge about the time series, e.g. the demand has weekly trend ($m = 7$) or the temperature has an annual trend ($m = 365$). In the absence of the domain knowledge, one can use Fourier transform to find the dominant frequencies of the time series and determine m based on that. We assume the parameter m is also known via the aforementioned or other processes.

B VARIABLE VARIANCE IN ARIMA

ARMA model assumes a fixed innovation variance of σ^2 as defined in this section. However, some time series do not fit into this criteria as the innovation variance might change from one point to the other. Auto Regressive Conditional Heteroskedasticity (ARCH) and its variations are a class of models that model the evolution of variance across samples (in contrast to ARMA model that captures the evolution of the mean across samples).

Definition 6 (GARCH(r,s)). *Given a time series* $\mathbf{x}_1^n$, *we model* x_i *for* $i > n$ *as*

$$\begin{aligned} x_i &\sim \mathcal{D}\left(\mu_i, \sigma_i^2\right) \\ \mu_i &= \mu \\ \sigma_i^2 &= \omega + \sum_{j=1}^{r} \alpha_j \sigma_{i-j}^2 - \sum_{k=1}^{s} \beta_k u_{i-k}^2 \\ u_i &= x_i - \mu_i \end{aligned} \tag{13}$$

where, α*'s,* β*'s and* ω *are the model parameters.*

Combining GARCH as defined above with ARMA as defined in (2) would allow us to capture the evolution of mean and variance of the distribution at the same time. These definitions can be extended to the ARIMA and seasonal cases. We skip including all of those definitions since it is trivial how to extend the current definitions to include those cases.

Definition 7 (ARMA(p,q)-GARCH(r,s)). *Given a* stationary *time series* $\mathbf{x}_1^n$, *we model* x_i *for* $i > n$ *as*

$$\begin{aligned} x_i &\sim \mathcal{D}\left(\mu_i, \sigma_i^2\right) \\ \mu_i &= \sum_{j=1}^{p} \phi_j \mu_{i-j} - \sum_{j=1}^{p} \phi_j u_{i-j} - \sum_{k=1}^{q} \theta_k u_{i-k} \\ \sigma_i^2 &= \omega + \sum_{j=1}^{r} \alpha_j \sigma_{i-j}^2 - \sum_{k=1}^{s} \beta_k u_{i-k}^2 \\ u_i &= x_i - \mu_i \end{aligned} \tag{14}$$

where, ϕ*'s,* θ*'s,* α*'s,* β*'s and* ω *are the model parameters.*