Analyzing and Forecasting All-India Tur (Arhar) Yields: A Time Series Approach

Dr. P. Sameerabanu

Assistant Professor, Department Of Mathematics, School Of Engineering And Technology, Dhanalakshmi Srinivasan University, Trichy - 621112 ,Tamilnadu, India.

Received:- 02 August 2024/ Revised:- 10 August 2024/ Accepted:- 16 August 2024/ Published: 31-08-2024

Abstract— *The agricultural sector plays a vital role in India's economy, with Tur (Arhar) being a significant crop. Accurate yield forecasting is essential for efficient agricultural planning and resource allocation. This study employs time series analysis and forecasting techniques to predict the all-India yield of lentils. Historical yield data were collected, preprocessed, and subjected to exploratory data analysis to identify trends and seasonal patterns. Various models, including ARIMA, SARIMA, and Exponential Smoothing, were evaluated for their forecasting performance. The models were trained on historical data and validated using metrics such as Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The results indicate that the chosen models provide reliable forecasts, which can aid policymakers and farmers in making informed decisions. The study highlights the importance of time series analysis in agricultural forecasting and provides a methodological framework for future research in crop yield prediction.*

Keywords— Tur(Arhar) Yield, AR, MA, ARMA, RMSE, MAE, SPSS Software.

I. INTRODUCTION

Lentils (masur) are a crucial pulse crop in India, contributing significantly to the country's food security and agricultural economy. As a major source of protein for a large segment of the population, lentils play an essential role in the dietary habits of millions of Indians. Given the importance of lentils, accurate forecasting of their yield is vital for effective agricultural planning, market stability, and policy formulation.

In the context of an ever-evolving climate and changing agricultural practices, predicting crop yields has become increasingly complex. Traditional methods of yield prediction often fall short in capturing the intricate patterns and trends present in agricultural data. This has led to the adoption of advanced statistical and machine learning techniques for time series analysis and forecasting.

Time series analysis offers a robust framework for analyzing historical yield data, identifying underlying patterns, and generating reliable forecasts. This study aims to apply various time series models to forecast the all-India yield of lentils. By leveraging historical yield data, the study seeks to develop models that can accurately predict future yields, thereby assisting policymakers, farmers, and stakeholders in making informed decisions.

The following sections detail the methodology used for data collection, preprocessing, exploratory data analysis, model selection, and evaluation. The study concludes with a discussion of the results and their implications for the agricultural sector in India.

Time series analysis and forecasting for agricultural yields, such as lentil (masur) in India, involve several steps. Here’s a general outline of the process:

1.1 Data Collection:

Gather historical yield data for lentils in India. This data can be obtained from government databases, agricultural research institutes, or international organizations like the Food and Agriculture Organization (FAO).

1.2 Data Preprocessing:

- **Cleaning**: Handle missing values, outliers, and any inconsistencies in the data.
- **Transformation**: Normalize or scale the data if necessary. You might also need to transform the data to make it stationary (e.g., using differencing).

1.3 Exploratory Data Analysis (EDA):

- **Trend Analysis**: Identify any long-term trends in the data.
- **Seasonal Analysis**: Determine if there are any seasonal patterns.
- **Plotting**: Visualize the data using line plots, histograms, and autocorrelation plots.

1.4 Model Selection:

Choose appropriate time series models. Common models include:

- **ARIMA (AutoRegressive Integrated Moving Average)**: Suitable for univariate time series data.
- **SARIMA (Seasonal ARIMA)**: Extension of ARIMA that handles seasonality.
- **Exponential Smoothing**: Simple models for short-term forecasting.
- **Machine Learning Models**: LSTM (Long Short-Term Memory) networks for more complex patterns.

1.5 Model Fitting:

- **Parameter Estimation**: Use techniques like grid search or auto ARIMA to find the best parameters for the chosen models.
- **Training**: Fit the model on historical data.

1.6 Model Evaluation:

- **Validation**: Split the data into training and test sets to evaluate the model's performance.
- **Metrics**: Use metrics like Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) to assess accuracy.

1.7 Forecasting:

- **Short-term vs Long-term**: Decide the forecasting horizon based on your needs.
- **Generate Forecasts**: Use the fitted model to predict future yields.

1.8 Post-Forecasting Analysis:

- **Interpretation**: Analyze the forecast results and interpret them in the context of agricultural planning.
- **Uncertainty Analysis**: Assess the confidence intervals and potential uncertainties in the forecasts.

1.9 Reporting and Visualization:

- **Visualization**: Plot the forecasted values along with historical data.
- **Reporting**: Prepare a report detailing the methodology, analysis, and forecasts.

II. LITERATURE REVIEW FOR ARIMA MODELS

- Advantages and limitations of different time-series models generally adopted have been critically reviewed. Further, comparison with reference to forecast accuracy of different models and their applications in the past are discussed below.
- The forecasting of energy consumption can be done based on Autoregressive Integrated Moving Average (ARIMA) models. There are other models namely used in forecasting energy consumption in economies. Multiple regression models, and artificial neural network models. There are four steps involved in this model and these steps have been explained by Ajith and Baikunth (2001) as model identification, parameter estimation, model diagnostics, and forecast verification and reasonableness.
- Zhu, Guo and Feng studied the issue of household energy consumption in China from the year 1980 to 2009 with construction VAR model. There two forecasting methods ARIMA and BVAR were used. The results showed that both of them can predict the sustained growth of Household Energy Consumption (HEC) trends.
- Ediger and Akar applied SARIMA (Seasonal ARIMA) methods to estimate the future primary fuel energy demand in Turkey from the year 2005 to 2020.
- Contreras and colleagues applied ARIMA methods to predict next day electricity price in mainland Spain and Californian markets. Conejo and colleagues applied wavelet transform and ARIMA models to predict day ahead electricity price of mainland Spain in the year 2002. Hence, the researcher performed a comparative study of ARIMA and ARMA models for a specific time series dataset.
- Box and Jenkins developed the autoregressive moving average to predict time series. There exists a vast literature for forecasting a univariate time series model based on neuro-fuzzy inference system. A fuzzy ARIMA model for forecasting foreign exchange market is presented.A hybrid ARIMA and neural network approach for forecasting time series is presented “Al-Fuhaid et al.” developed a neural network based short term load forecasting in Kuwait. “Che et al.” developed a hybrid model for forecasting short term electricity prices based on ARIMA and support vector regression. Different hybrid forecasting approaches are evaluated “Chengqun Yin et al.” forecasted short term load based on hybrid neural network model.

III. ANALYSIS PART

3.1 Sequence Plot:

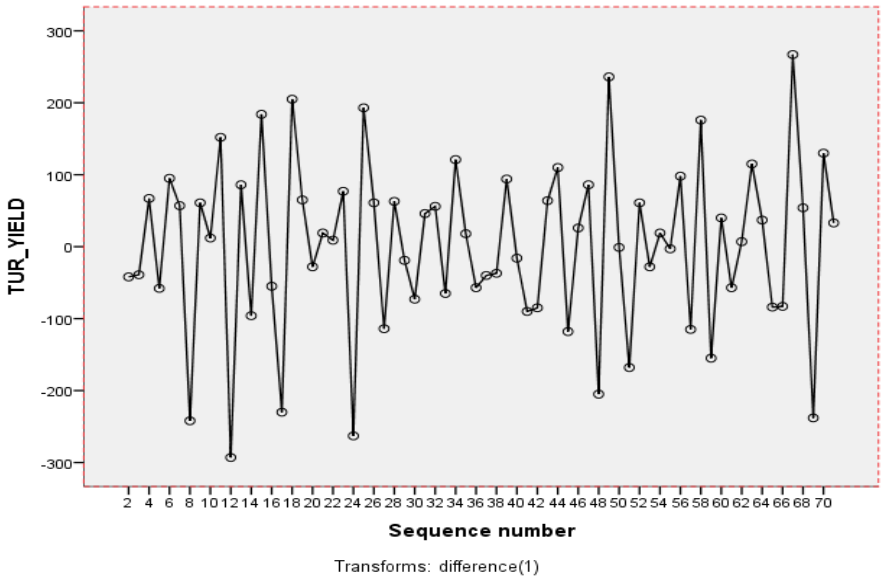

FIGURE 1: ACF and PACF

3.2 Identification:

The graphs of ACF and PACF are drawn for all the observed variables, to the fitted models.

TABLE 1
ACF AND PACF

	Autocorrelation	Std. Error[a]	Partial Autocorrelation	
1	-0.446	0.117	-0.446	0.12
2	-0.176	0.116	-0.468	0.12
3	0.121	0.115	-0.326	0.12
4	0.053	0.114	-0.207	0.12
5	0.03	0.114	-0.008	0.12
6	-0.078	0.113	0.021	0.12
7	0.044	0.112	0.114	0.12
8	-0.096	0.111	-0.077	0.12
9	0.128	0.11	0.024	0.12
10	-0.031	0.109	-0.005	0.12
11	-0.097	0.108	-0.091	0.12
12	0.145	0.107	0.074	0.12
13	-0.121	0.106	-0.054	0.12
14	0.057	0.105	0.006	0.12
15	0.036	0.104	0.067	0.12
16	-0.012	0.104	0.114	0.12

Assumed is independence (white noise)
chi-square approximation.

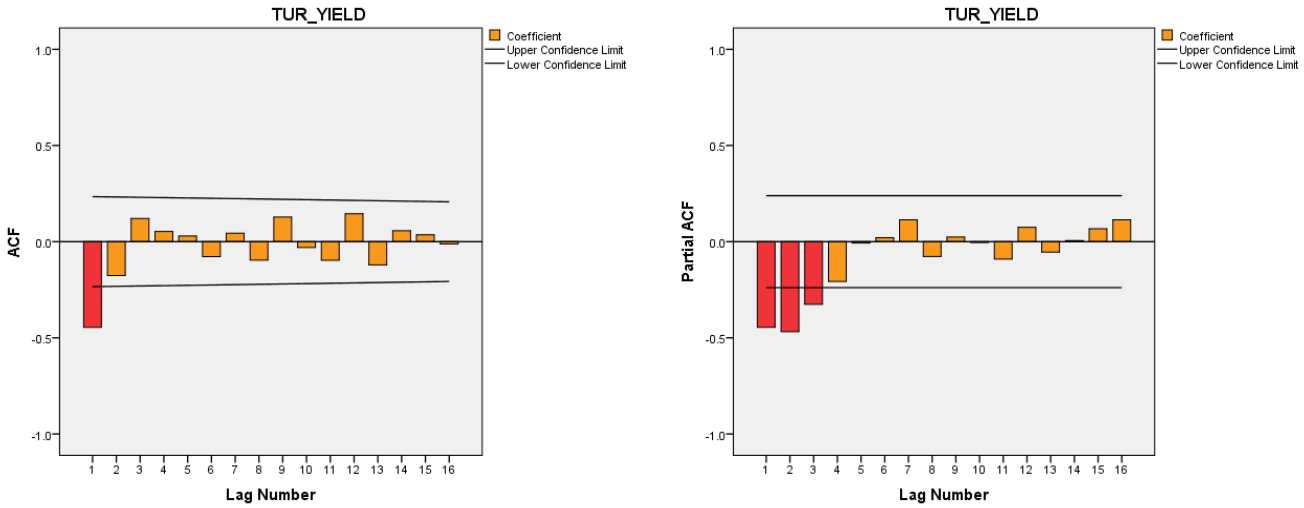

FIGURE 2: Plots of the ACF and the PACF for the yearly Tur Yield data difference series from 1950 to 2020

Figure 2 consists of plots of the ACF and the PACF for the yearly Tur Yield data difference series from 1950 to 2020, 95% confidence brands are plotted on the both panels. The ACF and the PACF of the difference values obtained by using the transformation $x_t = y_t - y_{t-1}$. The Box-Jenkins approach is applied to choose the value p and q by ACF and PACF plot. From PACF plot, it significantly spikes at lag1 and it could be viewed as dying out after lag1. It implies that an AR (1) model has to be build, while MA (1,2,3) from ACF plot. AIC of all the possible models are compared to find out a model to fit the data better than others having the lowest AIC value.

The ACF graph for gold price dies out slowly (exponentially decaying), with one spike in PACF that cuts after lag1,2,3. Data is stationary. Therefore, the first model for Tur Yield is initially identified as ARIMA (1, 1, 1), ARIMA (2, 1, 1), ARIMA (3,1,1).

3.3 Forecast:

In this section, ready defined models are used to forecast Tur Yield data from 1950 to 2020. The forecasting accuracy is also performed.

3.4 Forecasting:

Once the model adequacy is established, the series in question is forecasted for a specified period of time. It is always advisable to keep track of the forecast errors. Thus, depending on the magnitude of the errors, the model shall be re-evaluated. Therefore, in order to select the best ARIMA model, the best criteria has to be selected as mentioned below:

IV. METHODOLOGY

The criteria which have been used to make a fair comparison has been presented in this subsection. The framework comparison can be presented in more detail as follows:

The comparison of the performance of the models within two types of accuracy criteria have been adopted: Root Mean Squared Error (RMSE) and Mean Absolute Percentage Error (MAPE). Thus, these types of accuracy are illustrated as (Aggarwal et. al. (2008)):

4.1 For model:

For 95% confidence intervals, z(0.025) = 1.96

4.1.1 Time Series Modeler:

Model Description			Model Types		
Model ID	TUR_YIELD	Model_1	ARIMA(1,1,1)	ARIMA(2,1,1)	ARIMA (3,1,1)

4.1.2 Model Summary:

Fit Statistic	**ARIMA (1,1,1)**	**ARIMA (2,1,1)**	**ARIMA (3,1,1)**
	Mean	**Mean**	**Mean**
Stationary R-squared	0.424	0.462	0.467
RMSE	**91.394**	**89.014**	**89.231**
MAPE	10.525	10.219	10.276
MaxAPE	52.824	41.838	40.811
MAE	**72.366**	**70.2**	**70.57**
MaxAE	236.651	199.59	189.236
Normalized BIC	9.212	9.22	9.286

4.2 ARIMA (1,1,1):

Forecast:

Model		72	73	74	75	76	77	78	79	80	81
TUR_YIELD-Model_1	**Forecast**	821	828	829	830	831	832	833	835	836	837
	UCL	1003	1012	1017	1021	1026	1031	1036	1040	1045	1049
	LCL	639	644	641	638	636	634	631	629	627	625

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

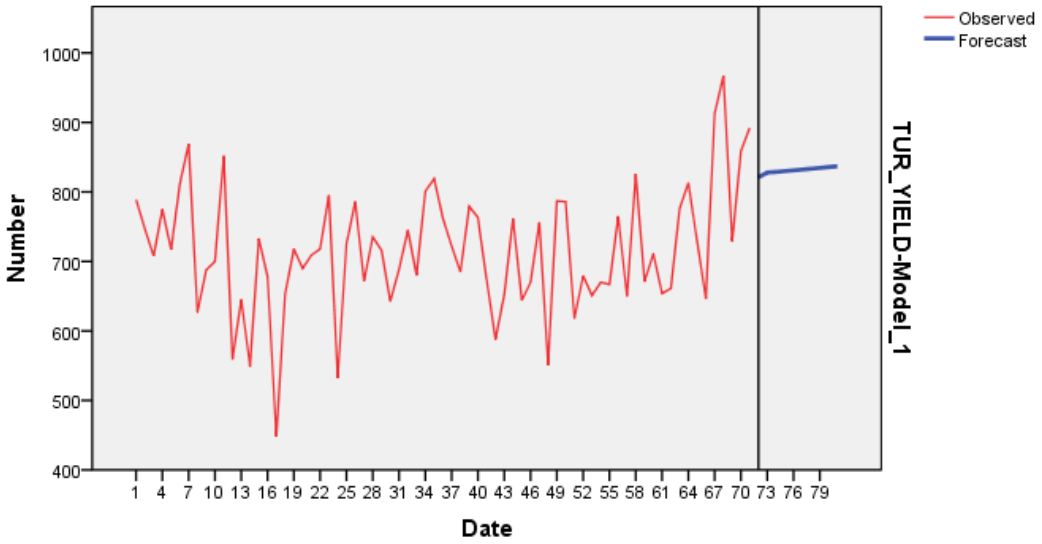

FIGURE 3: ARIMA (1,1,1) TUR_YIELD-Model_1 observed and Forecast

4.3 ARIMA (2,1,1):

ARIMA Model Parameters					Estimate	SE	t	Sig.
TUR_YIELD-Model_1	TUR_YIELD	No Transformation	Constant		1.322	2.839	0.466	0.643
			AR	Lag 1	-0.274	0.17	-1.614	0.111
				Lag 2	-0.32	0.145	-2.209	0.031
			Difference		1			
			MA	Lag 1	0.586	0.158	3.702	0

For Model:

$$Y_i = C_1 + \varphi (Y_{i-1}) + \theta \varepsilon_{1, i-1} \quad (1)$$

Where C is constant, ε_i is white noise

$$Y_i = Y_i - Y_{i-1} \quad (2)$$

Combine (1) and (2), we have:

$$Y_i - Y_{i-1} = C_1 + \varphi (Y_{i-1} - Y_{i-2}) + \varepsilon_1 + \theta_1 \varepsilon_{1, i-1} \quad (3)$$

Then,

$$Y_i = 1.322 - 0.274\, Y_{i-1} - 0.320\, Y_{i-2} + \varepsilon_1 + .586\, \varepsilon_{1, i-1} \quad (4)$$

Forecast:

Model		72	73	74	75	76	77	78	79	80	81
TUR_YIELD-Model_1	Forecast	827	836	857	850	847	852	854	854	856	857
	UCL	1004	1015	1036	1040	1045	1053	1060	1066	1072	1078
	LCL	649	657	677	660	650	651	648	642	639	636

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

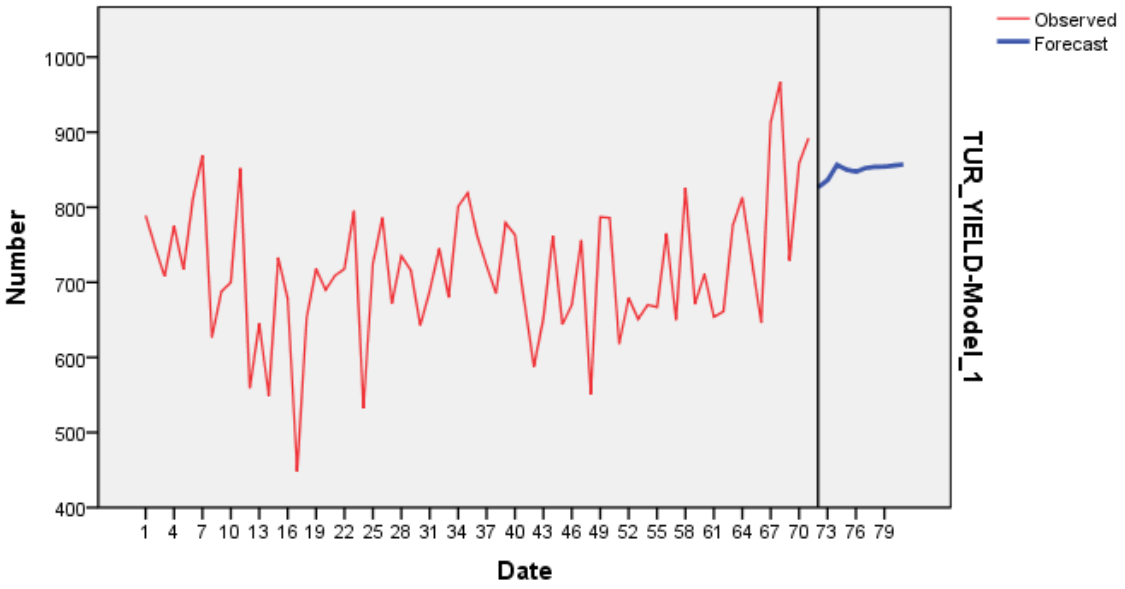

FIGURE 4: ARIMA (2,1,1) TUR_YIELD-Model_1 observed and Forecast

4.4 ARIMA(3,1,1):

Forecast:

Model		72	73	74	75	76	77	78	79	80	81
TUR_YIELD-Model_1	Forecast	851	836	859	865	857	857	863	865	864	866
	UCL	1029	1015	1039	1054	1058	1063	1073	1081	1087	1093
	LCL	673	657	680	676	657	652	653	648	642	638

For each model, forecasts start after the last non-missing in the range of the requested estimation period, and end at the last period for which non-missing values of all the predictors are available or at the end date of the requested forecast period, whichever is earlier.

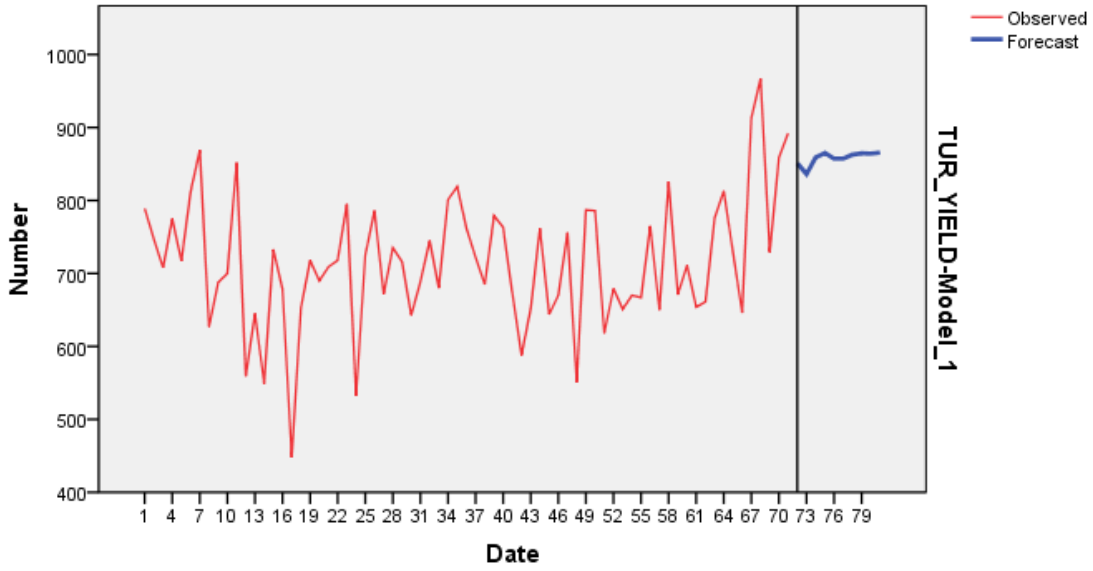

FIGURE 5: ARIMA (3,1,1) TUR_YIELD-Model_1 observed and Forecast

Finally it can be seen that ARIMA (2, 1, 1) provides a good fit for Rainfall data. It gives a fairly accurate forecasting. However, although forecasts from 1950 to 2020 are within the 95% percent interval, the graph shows that the red line of actual data has gradually moved out of the confidence interval .Such trend exactly coincides with the way of how the climate has evolved since great recession. But, the weakness of ARIMA model is that it could not predict such trend but rather assume the same pattern from 1950 to 2020, that is, the ARIMA model is not good for volatility analysis.

V. RESULT

TABLE 2
ACCURACY FORECAST FOR ARIMA MODELS

Statistical fit	ARIMA(1,1,1)	ARIMA(2,1,1)	ARIMA(3,1,1)
RMSE	91.394	89.014	89.231
MAPE	10.525	10.219	10.276
MAE	72.366	70.2	70.57

The best forecast is obtained from ARIMA (2,1,1) become it has low RMSE, MAPE value compared to other model.

REFERENCES

[1] Abdulhai, B., Porwal, H. and Recker, M. (1993). "Short-Term Freeway Traffic Flow prediction Using Genetically Optimised Time-Delay-Based Neural Networks". Presented at the 78th Annual Meeting of Transportation Research Board, Washington.

[2] Adenomon, MO., Ojehomon, VET. and Oyejola, BA. (2013). "Modelling the dynamic relationship between rainfall and temperature time series data in Niger State" Nigeria. Math Theory Model 3(4): 53–71.

[3] Adhikari, R., and Agrawal, R.K. (2008). "An Introductory Study on Time Series Modeling and Forecasting". Lap Lambert Academic Publishing.

[4] Agrawal, K., Ratnadip Adhikari, R. "An Introductory Study on Time Series Modeling and Forecasting". https:// arxiv.org/ ftp/ arxiv/ papers/ 1302/ 1302.6613.pdf.