Price relations wheat in major domestic and international markets

Samita Paudel

6/28/2019

1 Notes

- Formulate simple linear regression
- Multicollinearlity check...refer to Variance Inflation Factor
- Plot time series
- Autocorrelation check
- Drop variables with high multicollinarity
- Use

  - Darbin-Watson test for autocorrelation test (D-statistic), or
  - Brueuch Godfrey test for autocorrelation

- Use

  - Bursch-pagan test for heteroscadesticity
  - Or plot histogram

- In time series analysis,

  - Use regression with OLS
  - Plot data of each series (Fuel, prices, temperature)
  - Test dickey fueller test (Use no trend, no drift), use all possibility
  - Use augmented dickey fueller test if dickey fueller test does not capture the essense
  - Perform detrending with first order difference

2 Why domestic price is studied at district level

- Until recently, before federal structure of governance was into force, planning, budgeting, service delivery and policy interventions in Agriculture were all excised through district level bodies. Each district was itself accountable for market information accrual and reporting. Hence, most reliable form of price series data would be district level itself.

- Districts present isolated markets and well organized customer segments surrounding that. For e.g., food grain retail market of kathmandu is drastically different from that of kailali, because while in the former consumer segment has a larger role to play in determining of market demand, the market of farwestern terai region has significant share of producer segments in determining what and when to produce.
- District present different socio-economic narrative for food commodity trade which is heavily affected by the geographical context of the district itself. For e.g., Rupandehi market is more closely tied to bordering Indian market, because of minimal to none customs intervention in cross-country trade of food grain. The price effects of indian districts are more easily reflected in border region market prices, than in distant market such as Jumla and Surkhet.
- District markets information systems are more organized than local level markets, mostly because there are factors that buffer price volatility in district level. For e.g., government intervene through subsidized input supply, facilitated by district level agriculture offices when prices of agricultural inputs are hightened. Also service delivery system, for example that providing subsidized farmers loan, is carried out at district level.

3 Why only few domestic markets were selected for study

- These markets either have a large production volumes or strong consumer segment. Because districts of terai, and mostly those of Central to Farwestern region, show consistently high annual production volume (Terai is also dubbed grain basket of Nepal) major producer districts in terai – Kailali, Rupandehi, Parsa are included in the study. At the same time Chitwan and Kathmandu districts have prominance of consumers.
- Some of the features that justify suitability of inclusion of abovementioned districts are presented:

  – Kailali district lies in farwestern region. It borders with India through Uttar Pradesh state. The district has huge chunks of land annually allocated to Wheat production (How much in total ???, what percentage of total wheat cultivated area ???). The farmer segment comprises mainly of Tharu community.
  – Rupandehi district is located in Western terai region. The district adjoining with Indian market of Uttar Pradesh state, likewise, has large volume of grain production arising from Wheat cultivation.
  – Parsa district lies in Central terai part of Nepal. It border with India through Bihar state and the district cultivates Wheat in large volume (areawise how much ???).

- Kathmandu and Chitwan markets are mostly dominanted by consumer segment, hence the retail price series of these districts are expected to differ from that of producer market districts.
- ???Some of the food market features of Kathmandu and Surkhet districts???

- Treatment of isolated markets into arbitrary category of consumer and producer districts will provide a more complete picture of the situation of region they represent – Farwestern terai, Western-central terai and Eastern-central terai regions, respectively. This form of classification is expected to improve interpretability and overall increase predictive accuray of model in the face of price shocks.

3.1 Study districts and market centres
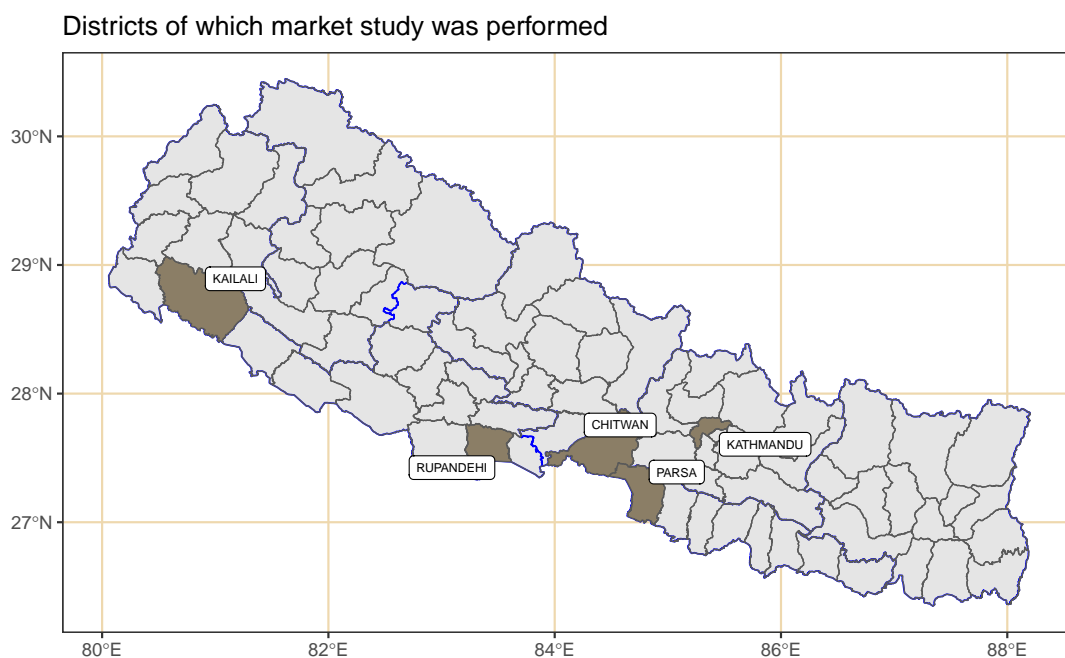
A map of study districts.

Figure 1: Geographical context of selected district markets

4 Study variables

4.1 Retail price of wheat in major domestic markets

Price series of domestic markets were selected for study. The data represent imbalanced series of following 5 districts: Kailali, Rupandehi, Parsa, Kathmandu, Chitwan.

4.2 Precipitation of Nepalese district markets

4.3 Foreign exchange
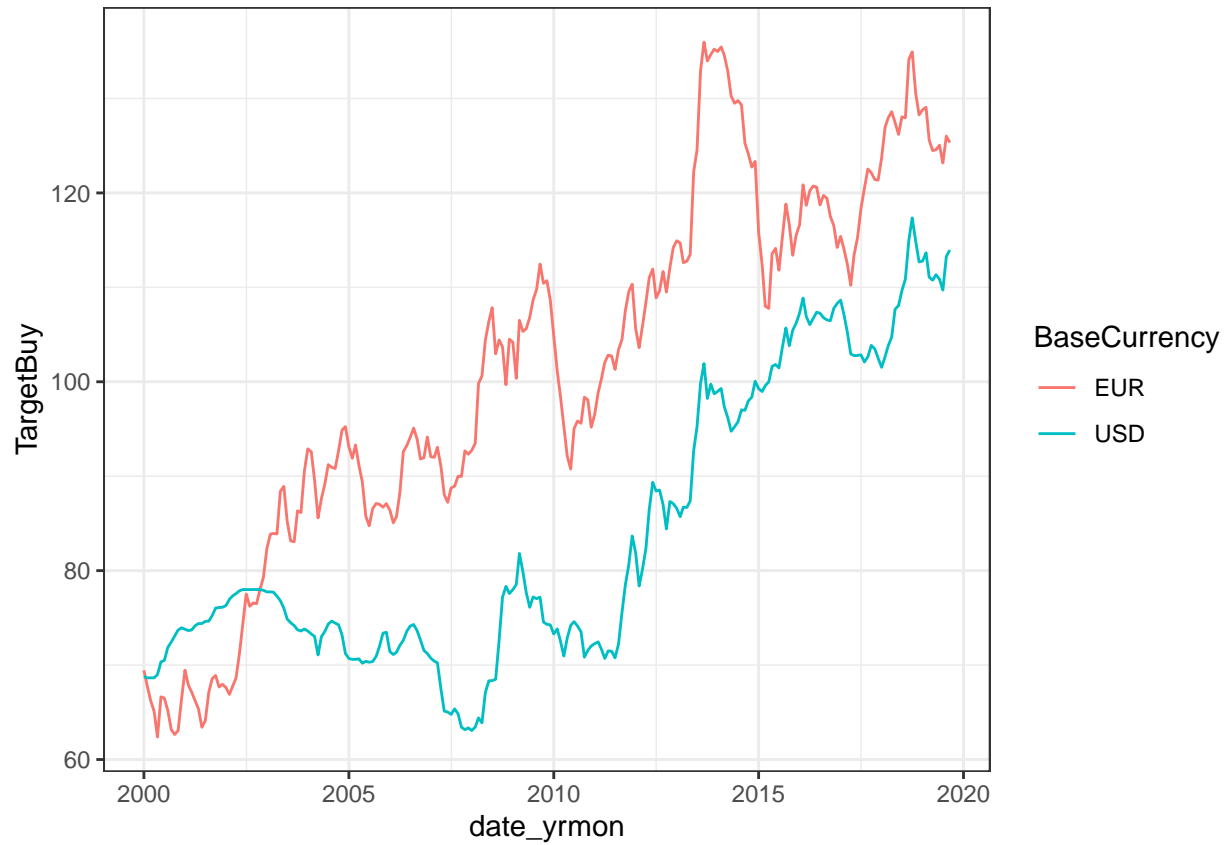
4.4   Fuel prices

4.5   International and India's bordering region prices
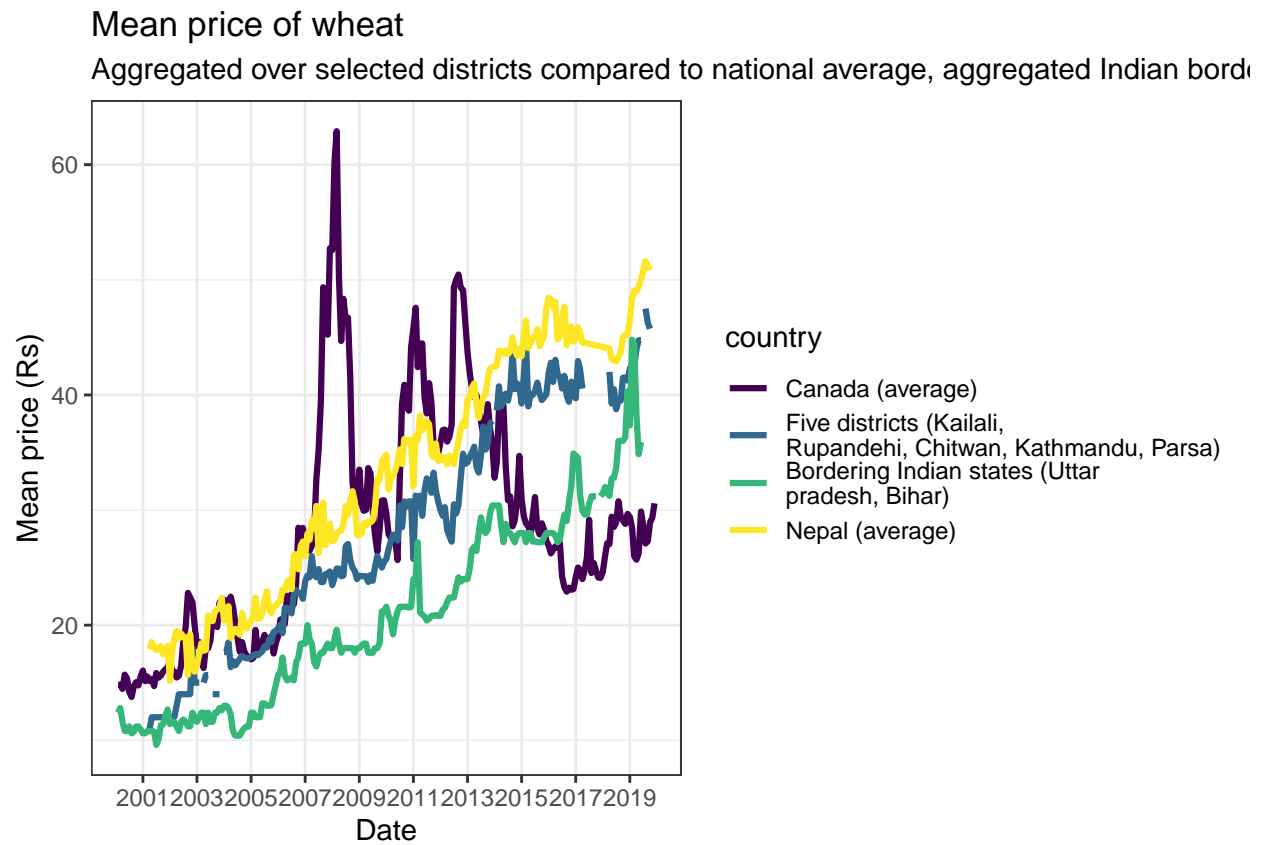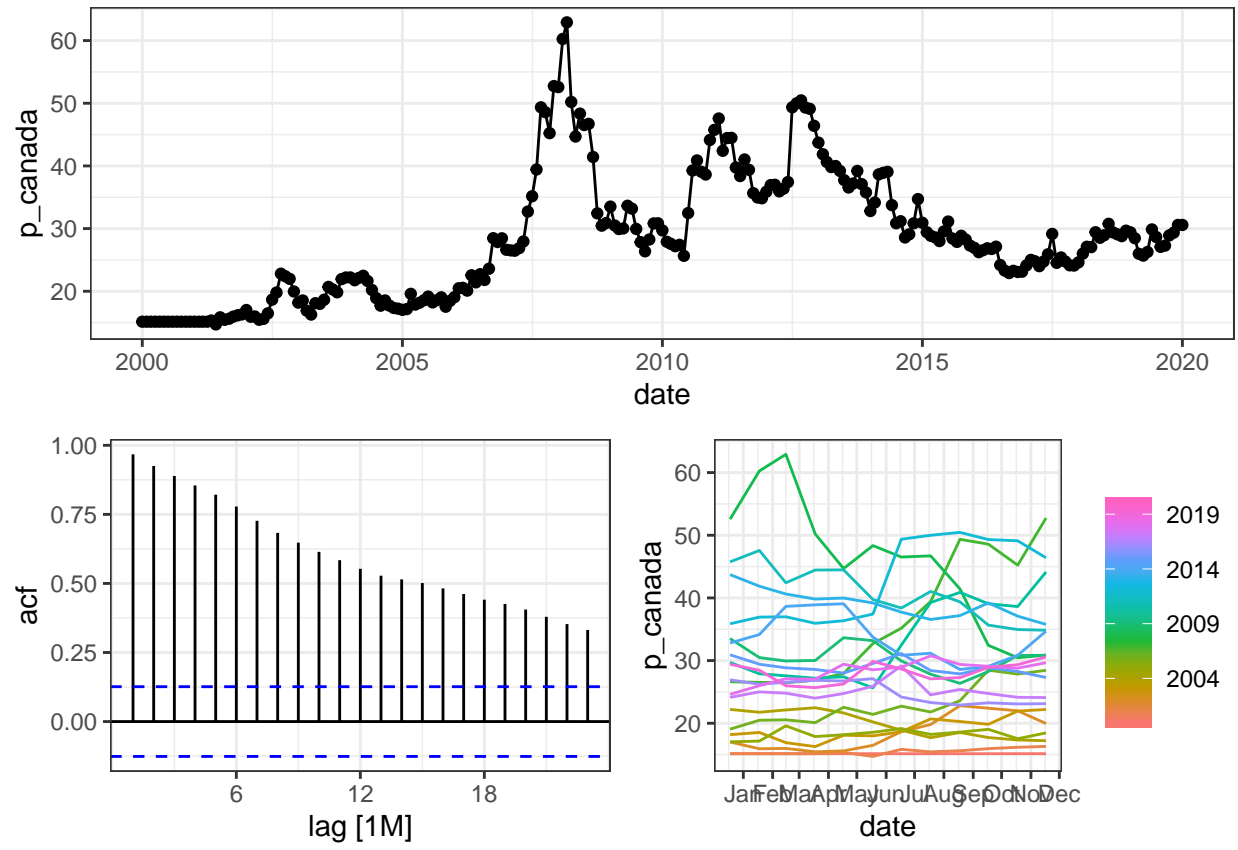
4.6   Cost of production
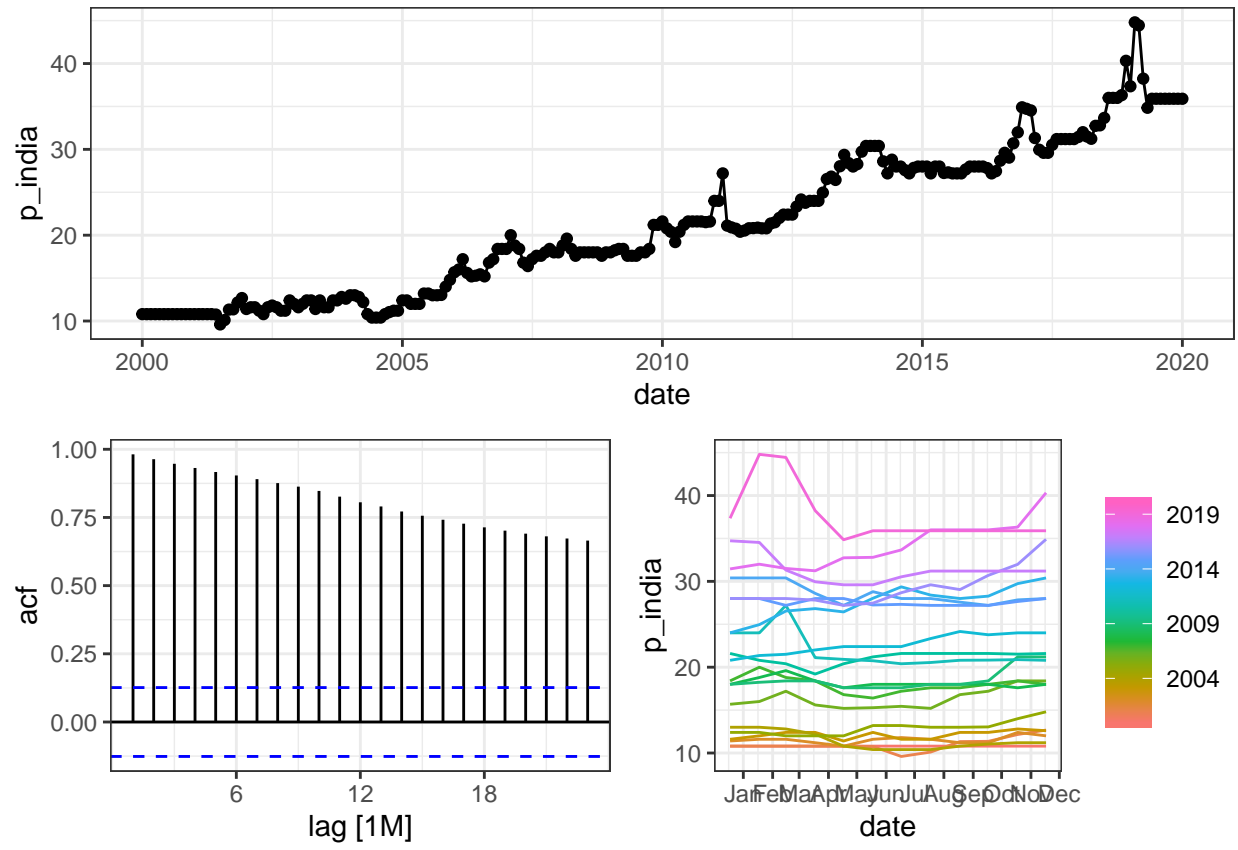

4.7   Wholesale price

4.8   Combined series

4.9   Descriptive time series summary

Plot of wheat retail price series of nepal and international (canada and bordering indian region) markets, in comparison to national average price.

Mean price of wheat

Aggregated over selected districts compared to national average, aggregated Indian borde
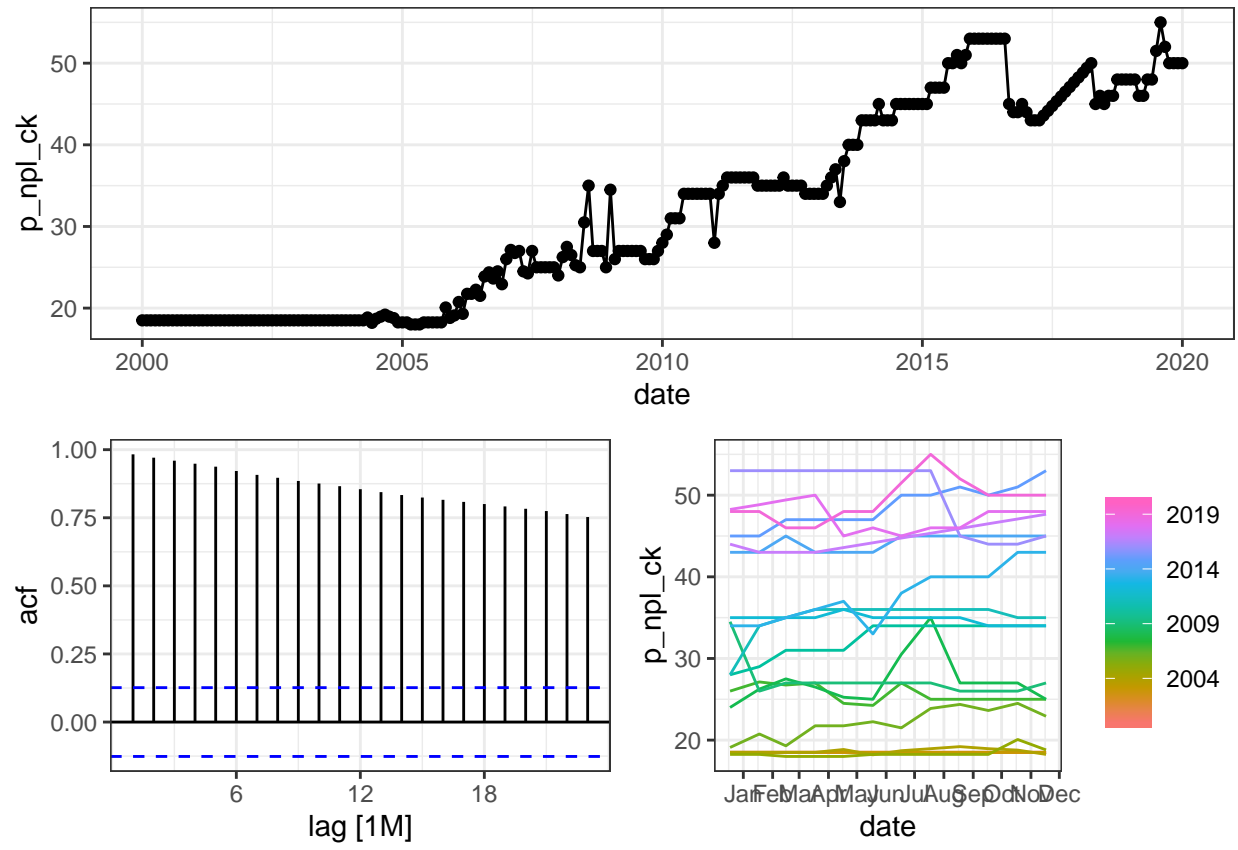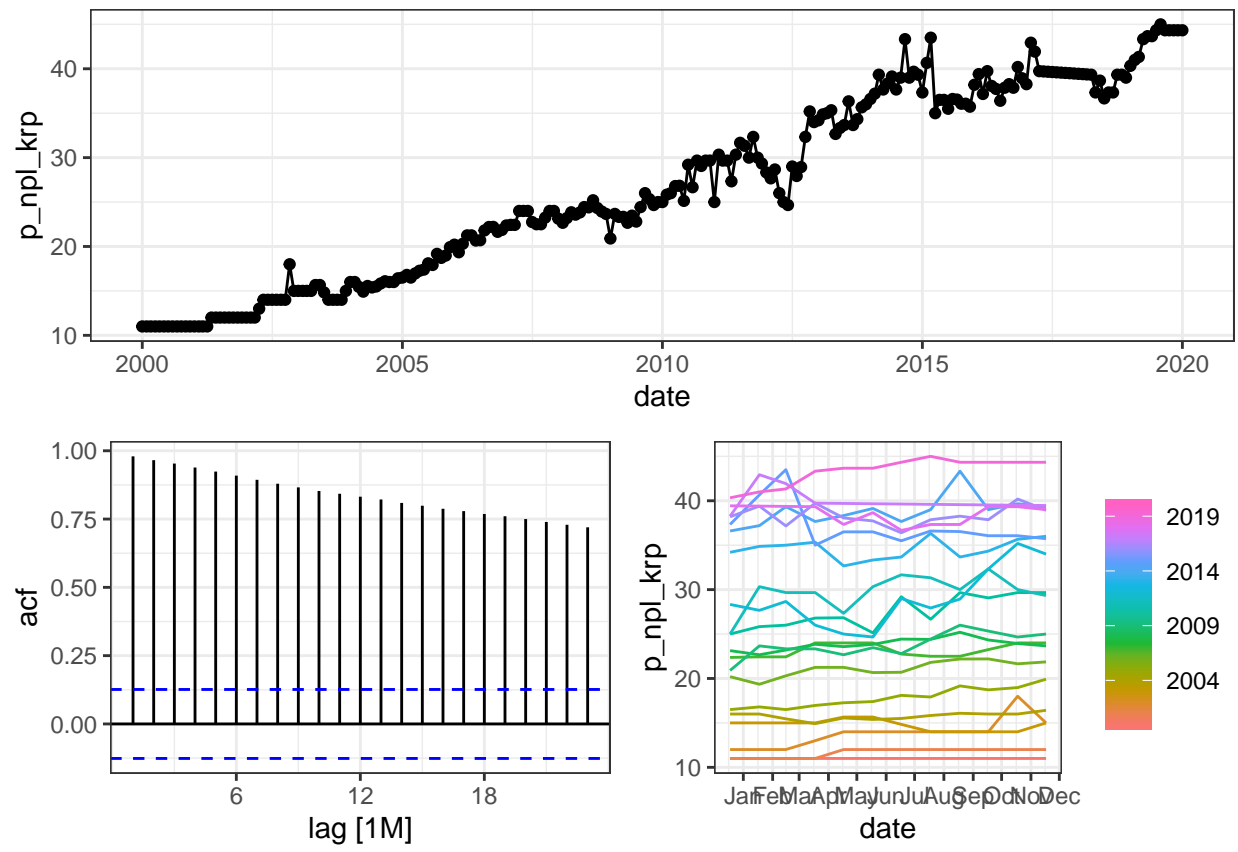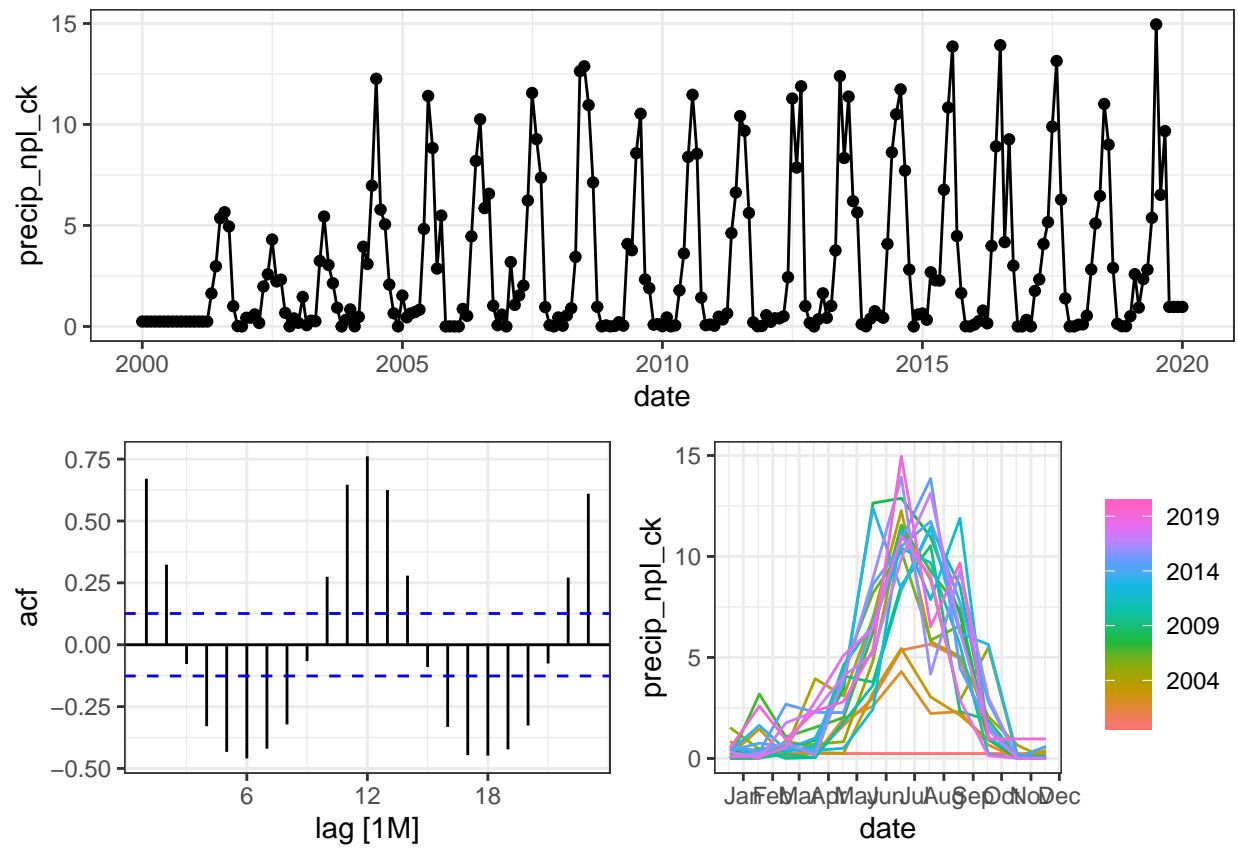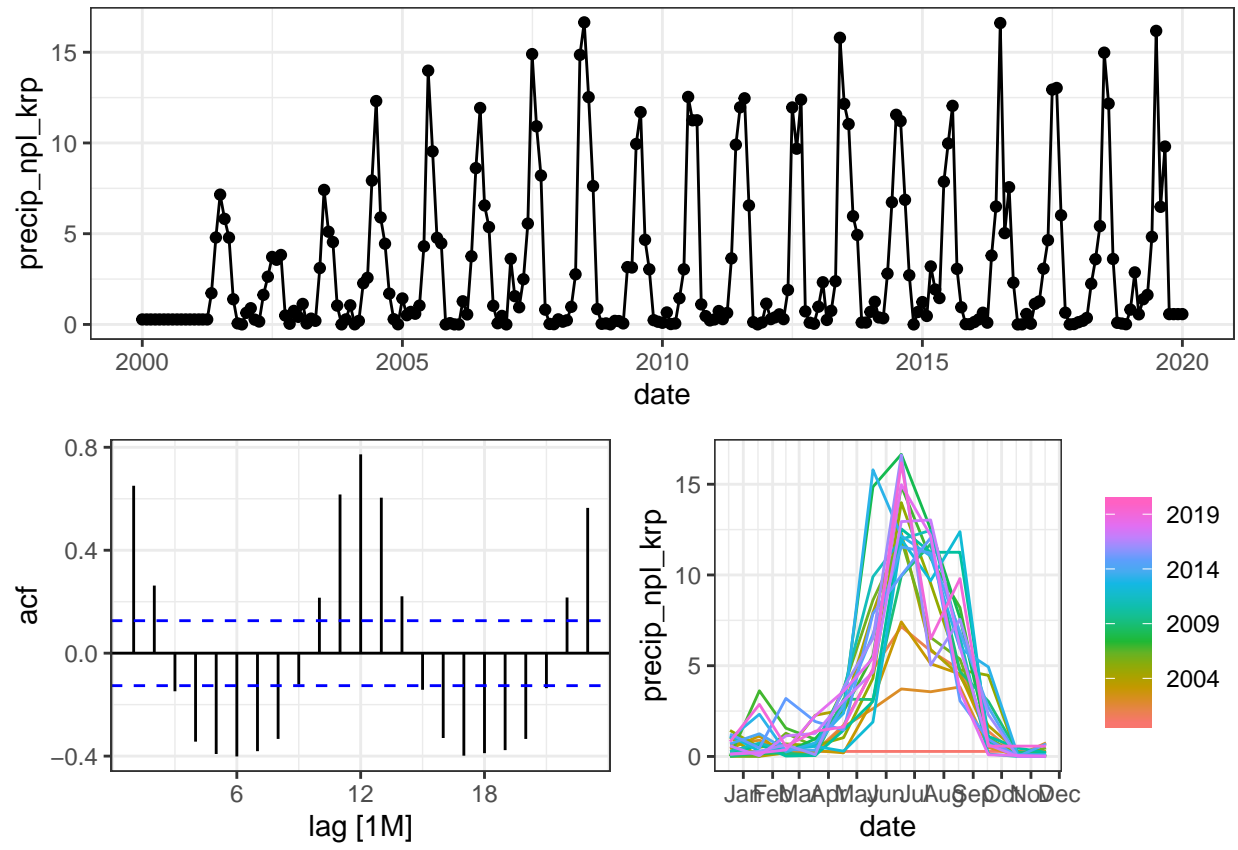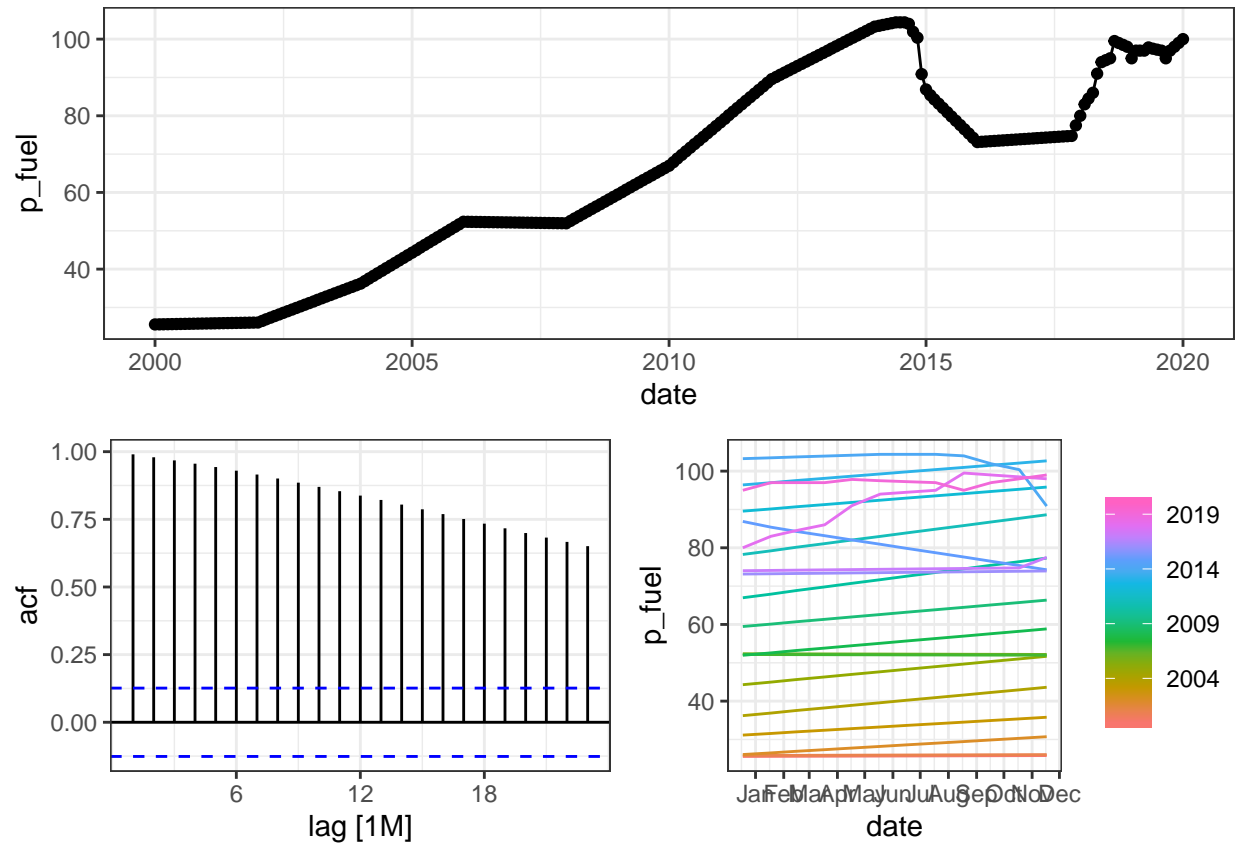
Time series plot of wheat retail price series is presented in above figures with some diagnostic plot alongside. The lineplot shows that there exist some time gaps at random periods. Autocorrelation of consecutive first order differenced lags is similarly shown.

5   Linear regression model formulation for price series

Table 1: ANOVA of regression between log(price) of producer districts (as dependent variable) and 6 regressor variables (log(price) of wheat in international market, log(price) of wheat in indian bordering states, log(price) of wheat in consumer districts, precipitation of consumer districts, precipitation of producer districts and log(price) of fuel)

term	df	sumsq	meansq	statistic	p.value
lp canada	1	18.986	18.986	2957.318	0.000
lp india	1	23.281	23.281	3626.404	0.000
lp npl ck	1	0.270	0.270	42.071	0.000
precip npl ck	1	0.060	0.060	9.376	0.002
precip npl krp	1	0.023	0.023	3.567	0.060
lp fuel	1	0.741	0.741	115.371	0.000
Residuals	234	1.502	0.006	NA	NA

Table 2: ANOVA of regression between log(price) of consumer districts (as dependent variable) and 6 regressor variables (log(price) of wheat in international market, log(price) of wheat in indian bordering states, log(price) of wheat in producer districts, precipitation of consumer districts, precipitation of producer districts and log(price) of fuel)

term	df	sumsq	meansq	statistic	p.value
lp canada	1	11.312	11.312	1076.770	0.000
lp india	1	21.523	21.523	2048.625	0.000
lp npl krp	1	0.298	0.298	28.335	0.000
precip npl ck	1	0.035	0.035	3.363	0.068
precip npl krp	1	0.019	0.019	1.779	0.184
lp fuel	1	0.051	0.051	4.877	0.028
Residuals	234	2.458	0.011	NA	NA

The regression above (Table 2 and 1) is obtained on fitting the model Equation 1.

$$
\begin{aligned}
price_{\text{producer districts}} = {} & price_{\text{canada}} + price_{\text{indian bordering states}} \\
& + price_{\text{consumer districts}} + precipitation_{\text{producer districts}} \\
& + precipitation_{\text{consumer districts}} + fuel\ price_{\text{national average}}
\end{aligned}
\tag{1}
$$

This presents a typical case of spurious relationship among variables, where variables having times series attributes show exceptionally high association among them. For example, all three aggregated wheat price series we consider in the regression (price in Canada, price in

India, and price in Nepalese consumer markets) which show highly significant association. This is problematic and misleading, because without accounting for linear time trend they show very unstable variance, which increases at extremes (more recent time period). The phenomena of possible presence of heteroskedasticity is shown in the residual-vs-fit plot in Figure 2.

Producer districts

Figure 2: Residual (pearsons') versus fit plot of the linear regression without accounting for time attributes of the series

Simply after incorporating linear trend of time in the regression model (Equation 1), we get a different statistic.

$$
\begin{aligned}
price_{\text{producer districts}} = {} & date + price_{\text{canada}} + price_{\text{indian bordering states}} \\
& + price_{\text{consumer districts}} + precipitation_{\text{producer districts}} \\
& + precipitation_{\text{consumer districts}} + fuel\ price_{\text{national average}}
\end{aligned}
\tag{2}
$$

However, even then there is likely presence of biased variance, as shown in Figure 5 which uses Equation 2 for model fitting.

Consumer districts
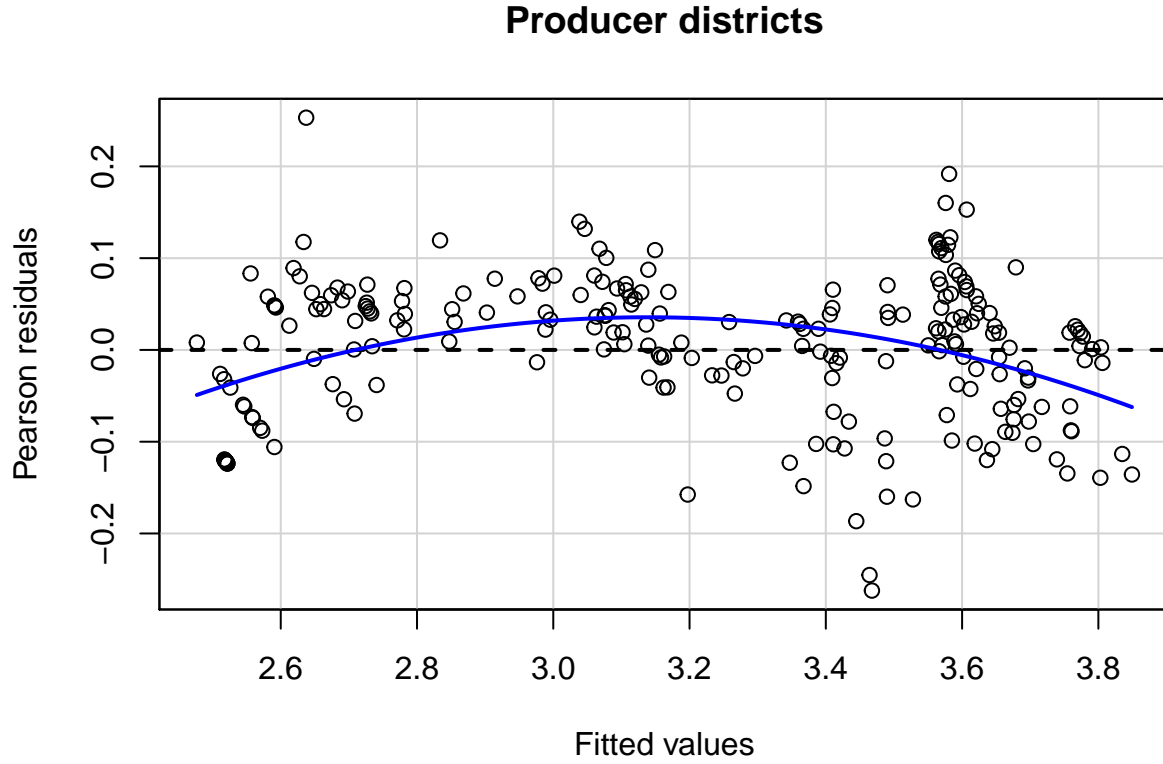

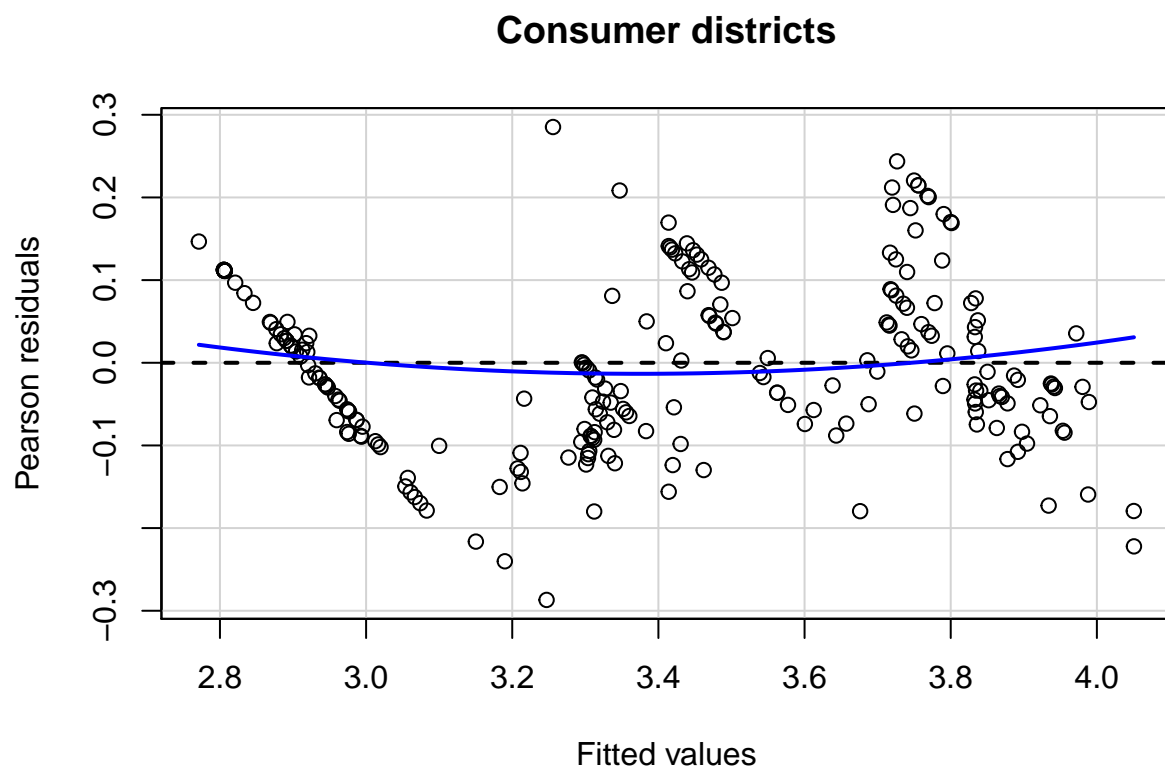Figure 3: Residual (pearsons') versus fit plot of the linear regression without accounting for time attributes of the series
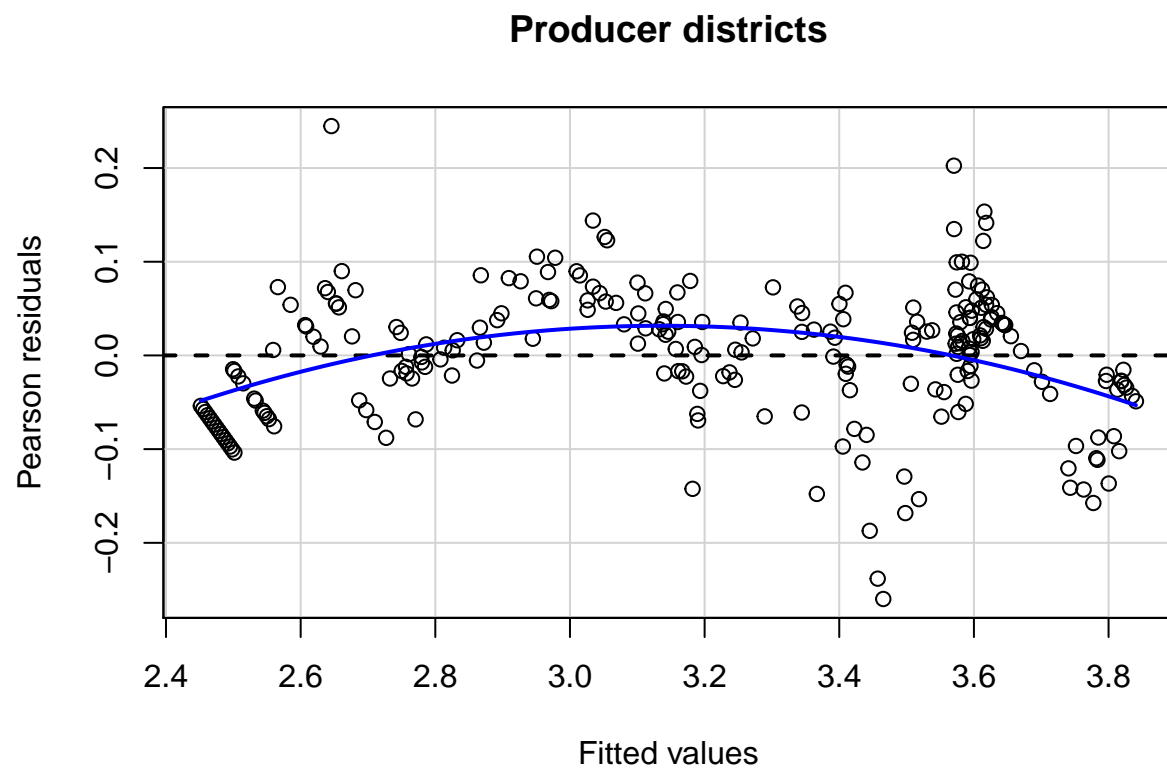
Producer districts

Figure 4: Residual (pearsons') versus fit plot of the linear regression after incorporating linear time trend of the price series
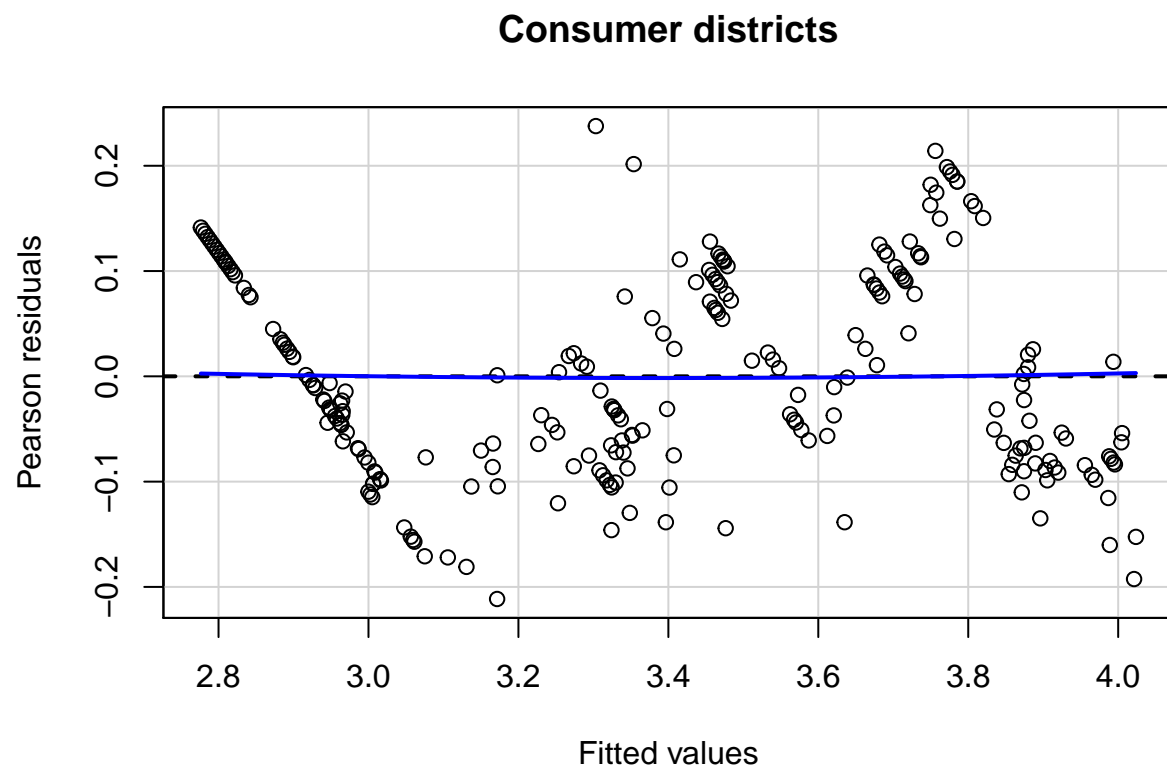
Figure 5: Residual (pearsons') versus fit plot of the linear regression after incorporating linear time trend of the price series

Table 3: Breusch-Pagan test for heteroskedasticity of price series, modeled by the regression Equation **??**eqn:lm2) for two domestic price series (producers districts and consumer districts)

statistic	p.value	parameter	method
4.477515	0.7234231	7	studentized Breusch-Pagan test
8.230593	0.3127004	7	studentized Breusch-Pagan test

5.1 Testing for heteroskedasticity

Below we test the linear model (Equation 2 and its counterpart with consumer districts as dependent variable and same independent variables with linear time trend) for presence of heteroskedasticity using Breusch-Pagan test. The test fits a linear regression model to the residuals of a linear regression model (by default the same explanatory variables are taken as in the main regression model) and rejects if too much of the variance is explained by the additional explanatory variables.

Breusch pagan test uses the null hypothesis of Homoscadesticity while testing studentized residuals. Hence, if the null hypothesis is rejected ($p < 0.05$), there is possible presence of heteroskedasticity.

A possible measure to removing non-stationary trend in the series is by differencing (with `diff`). However, before progressing we confirm that justifiable lag operations can infact render the series free of trends. For this, two popular unit test routines are performed – Augmented Dickey-Fueller test and KPSS test.

6 Unit root testing

6.1 Unit root (ADF and KPSS) test of retail price

The ADF, available in the function `adf.test()` (in the package `tseries`) implements the t-test of $H_0 : \gamma = 0$ in the regression, below.

$$\Delta Y_t = \beta_1 + \beta_2 t + \gamma Y_{t-1} + \sum_{i=1}^{m} \delta_i \Delta Y_{t-i} + \varepsilon_t \tag{3}$$

The null is therefore that x has a unit root. If only x has a non-unit root, then the x is stationary (rejection of null hypothesis).

The ADF test was parametrized with the alternative hypothesis of stationarity. This extends to following assumption in the model parameters;

$$-2 \leq \gamma \leq 0 \text{ or } (-1 < 1 + \phi < 1)$$

Table 4: Unit root test of log(price) series variables (both dependent and independent)

series	test	lprice pvalue	lprice tstatistic	lprice null accepted
lp canada	adf	0.45	-2.29	TRUE
lp india	adf	0.08	-3.23	TRUE
lp npl ck	adf	0.27	-2.73	TRUE
lp npl krp	adf	0.48	-2.22	TRUE
lp fuel	adf	0.64	-1.86	TRUE
lp canada	kpss	0.01	2.16	FALSE
lp india	kpss	0.01	4.79	FALSE
lp npl ck	kpss	0.01	4.80	FALSE
lp npl krp	kpss	0.01	4.74	FALSE
lp fuel	kpss	0.01	4.28	FALSE

`k` in the function refers to the number of δ lags, i.e., $1, 2, 3, ..., m$ in the model equation.

The number of lags `k` defaults to `trunc((length(x)-1)^(1/3))`, where `x` is the series being tested. The default value of `k` corresponds to the suggested upper bound on the rate at which the number of lags, `k`, should be made to grow with the sample size for the general ARMA(p,q) setup `citation(package = "tseries")`.

For a Dickey-Fueller test, so only up to AR(1) time dependency in our stationary process, we set `k = 0`. Hence we have no δs (lags) in our test.

The DF model can be written as:

$$Y_t = \beta_1 + \beta_2 t + \phi Y_{t-1} + \varepsilon_t$$

It can be re-written so we can do a linear regression of ΔY_t against t and Y_{t-1} and test if ϕ is different from 0. If only, ϕ is not zero and assumption above $(-1 < 1 + \phi < 1)$ holds, the process is stationary. If ϕ is straight up 0, then we have a random walk process – all white noise.

$$\Delta Y_t = \beta_1 + \beta_2 t + \gamma Y_{t-1} + \varepsilon_t$$

Alternative to above discussed tests, the Phillips-Perron test with its nonparametric correction for autocorrelation (essentially employing a HAC estimate of the long-run variance in a Dickey-Fuller-type test instead of parametric decorrelation) may be used. It is available in the function `pp.test()`.

6.2   Unit root test based lag order differencing determination

An alternative to decomposition for removing trends is differencing (Woodward, Gray, and Elliott 2017). We define the difference operator as,

$$\nabla x_t = x_t - x_{t-1}, \tag{4}$$

and, more generally, for order d

$$\nabla^d x_t = (1 - \mathbf{B})^d x_t, \tag{5}$$

Where $\mathbf{B}$ is the backshift operator (i.e., $\mathbf{B}^k x_t = x_{t-k}$ for $k \geq 1$).

Applying the difference to a random walk, the most simple and widely used time series model, will yield a time series of Gaussian white noise errors $\{w_t\}$:

$$\begin{aligned} \nabla(x_t &= x_{t-1} + w_t) \\ x_t - x_{t-1} &= x_{t-1} - x_{t-1} + w_t \\ x_t - x_{t-1} &= w_t \end{aligned} \tag{6}$$

We use an implementation of time series differencing based on optimal lag length in order to render series stationary. The first lag order differenced log(price) series (derived based on "kpss" test statistic) is shown in Figure 6. The kpss test, however, determined the precipitation series to be integrated of order null. This is due to insensitivity of test (model) to seasonal lag component, which is infact nicely captured by addtive or multiplicative trend decomposition (discussed ahead).

The first order differencing of price renders series stationary. However, precipitation series shows distinct components. Here, a linear model is fitted to the trend component, decomposed (into seasonal and trend components) from two precipitation series (producer and consumer districts), along with other terms of Equation 2 and the ANOVA output is presented.

6.2.1   Detrending precipitation

Here linear model is fitted with trend component only that of precipitation.

7   VAR model

7.1   VAR

VAR is a system regression model, i.e., there are more than one dependent variable. The regression is defined by a set of linear dynamic equations where each variable is specified as a function of an equal number of lags of itself and all other variables in the system. Any additional variable, adds to the modeling complexity by increasing an extra equation to be estimated.
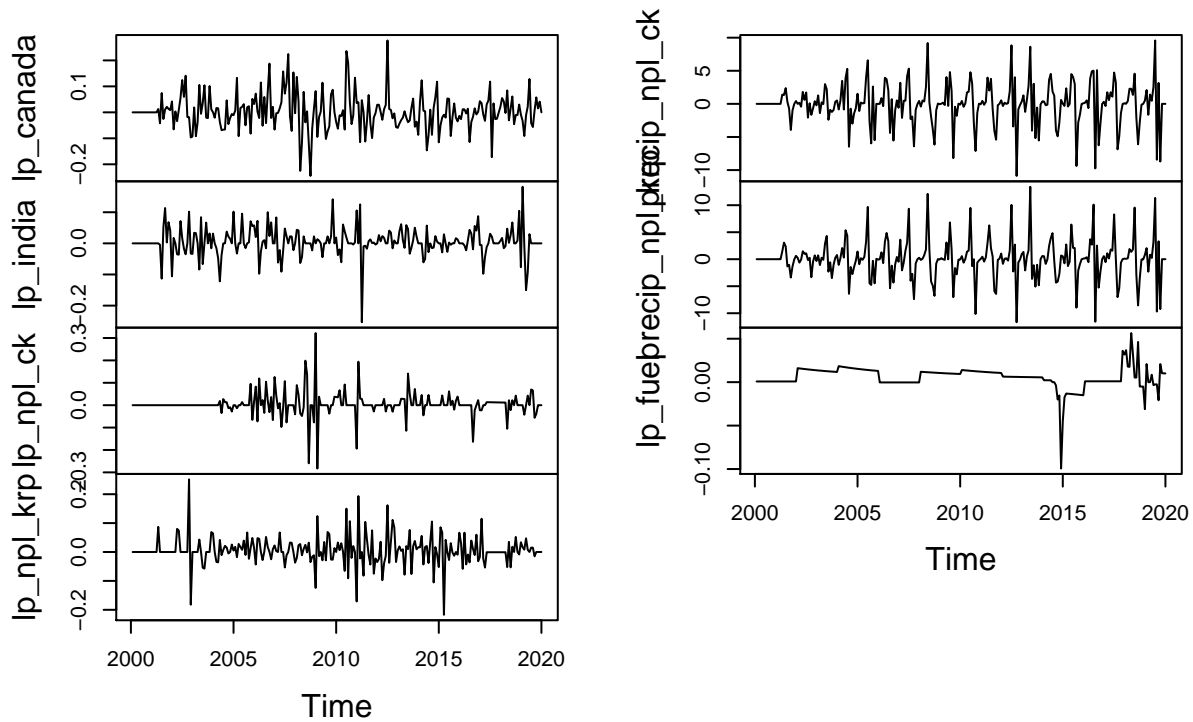
Figure 6: Plot of differenced time series for various lag length determined by kpss statistic.

Table 5: ANOVA of linear model fit with trend component of precipitation series with price(producer market) as dependent variable.

term	df	sumsq	meansq	statistic	p.value
date	1	36.65	36.65	8847.11	0.00
lp_canada	1	0.72	0.72	174.54	0.00
lp_india	1	0.04	0.04	8.67	0.00
lp_npl_ck	1	0.01	0.01	1.55	0.21
precip_npl_ck_trend	1	0.39	0.39	93.38	0.00
precip_npl_krp_trend	1	0.00	0.00	0.01	0.91
lp_fuel	1	0.15	0.15	37.18	0.00
Residuals	221	0.92	0.00	NA	NA

Table 6: ANOVA of linear model fit with trend component of precipitation series with price(consumer market) as dependent variable.

term	df	sumsq	meansq	statistic	p.value
date	1	30.29	30.29	3966.52	0.00
lp_canada	1	0.06	0.06	7.90	0.01
lp_india	1	0.09	0.09	11.81	0.00
lp_npl_krp	1	0.01	0.01	1.23	0.27
precip_npl_ck_trend	1	0.14	0.14	17.72	0.00
precip_npl_krp_trend	1	0.27	0.27	34.74	0.00
lp_fuel	1	0.03	0.03	4.08	0.04
Residuals	221	1.69	0.01	NA	NA

The vector autoregression (VAR) model extends the idea of univariate autoregression to k time series regressions, where the lagged values of *all k* series appear as regressors. Put differently, in a VAR model we regress a *vector* of time series variables on lagged vectors of these variables. As for $AR(p)$ models, the lag order is denoted by p so the $VAR(p)$ model of two variables X_t and Y_t $(k = 2)$ is given by a vector of equations (Equation 7).

$$
\begin{aligned}
Y_t &= \beta_{10} + \beta_{11}Y_{t-1} + \cdots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \cdots + \gamma_{1p}X_{t-p} + u_{1t}, \\
X_t &= \beta_{20} + \beta_{21}Y_{t-1} + \cdots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \cdots + \gamma_{2p}X_{t-p} + u_{2t}.
\end{aligned}
\tag{7}
$$

The βs and γs can be estimated using OLS on each equation.

Simplifying this to a bivariate $VAR(1)$, we can write the model in matrix form as:

$$
Y_t = \beta_0 + \beta_1 Y_{t-1} + \mu_t
\tag{8}
$$

Where,

- Y_t, Y_{t-1} and μ_t are (2 x 1) column vectors
- β_0 is a (2 x 1) column vector
- β_1 is a (2 x 2) matrix

also,

$$
Y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}, \ Y_{t-1} = \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix}
$$

$$
\mu_t = \begin{pmatrix} \mu_{1t} \\ \mu_{2t} \end{pmatrix}, \beta_0 = \begin{pmatrix} \beta_{10} \\ \beta_{20} \end{pmatrix}, \beta_1 = \begin{pmatrix} \beta_{11} & \alpha_{11} \\ \alpha_{21} & \beta_{21} \end{pmatrix}
$$

It is straightforward to estimate VAR models in `R`. A feasible approach is to simply use `lm()` for estimation of the individual equations. Furthermore, the `vars` package provides standard tools for estimation, diagnostic testing and prediction using this type of models.

Only when the assumptions presented below hold, the OLS estimators of the VAR coefficients are consistent and jointly normal in large samples so that the usual inferential methods such as confidence intervals and *t*-statistics can be used (Metcalfe and Cowpertwait 2009).

Two series $w_{x,t}$ and $w_{y,t}$ are bivariate white noise if they are stationary and their cross-covariances $\gamma_{xy}(k) = Cov(w_{x,t}, w_{y,t+k})$ satisfies

$$\gamma_{xx}(k) = \gamma_{yy}(k) = \gamma_{xy}(k) = 0 \text{ for all } k \neq 0$$

The parameters of a var(p) model can be estimated using the `ar` function in `R`, which selects a best-fitting order p based on the smallest information criterion values.

The structure of VARs also allows to jointly test restrictions across multiple equations. For instance, it may be of interest to test whether the coefficients on all regressors of the lag p are zero. This corresponds to testing the null that the lag order $p-1$ is correct. Large sample joint normality of the coefficient estimates is convenient because it implies that we may simply use an *F*-test for this testing problem. The explicit formula for such a test statistic is rather complicated but fortunately such computations are easily done using the `ttcode("R")` functions we work with in this chapter. Just as in the case of a single equation, for a multiple equation model we choose the specification which has the smallest $BIC(p)$, where

$$BIC(p) = \log\left[\det(\widehat{\Sigma}_u)\right] + k(kp+1)\frac{\log(T)}{T}.$$

with $\widehat{\Sigma}_u$ denoting the estimate of the $k \times k$ covariance matrix of the VAR errors and $\det(\cdot)$ denotes the determinant.

As for univariate distributed lag models, one should think carefully about variables to include in a VAR, as adding unrelated variables reduces the forecast accuracy by increasing the estimation error. This is particularly important because the number of parameters to be estimated grows qudratically to the number of variables modeled by the VAR.

Table 7: Model performance indicators of VAR(AR(1)) model for wheat log(price) series with two domestic (producer and consumer) markets as endogeneous variables and other price series, precipitation and fuel series as exogeneous regressors.

log_lik	AIC	AICc	BIC
727.44	-1414.88	-1410.822	-1346.293

Table 8: Model coefficients of VAR(AR(1)) model for for wheat log(price) series with two domestic (producer and consumer) markets as endogeneous variables and other price series, precipitation and fuel series as exogeneous regressors.

term	.response	estimate	std.error	statistic	p.value
lag(lp_npl_ck,1)	lp_npl_ck	0.843	0.033	25.449	0.000
lag(lp_npl_krp,1)	lp_npl_ck	0.066	0.043	1.534	0.127
constant	lp_npl_ck	0.105	0.045	2.328	0.021
lp_canada	lp_npl_ck	-0.003	0.015	-0.226	0.821
lp_india	lp_npl_ck	0.094	0.038	2.468	0.014
precip_npl_ck	lp_npl_ck	-0.002	0.004	-0.501	0.617
precip_npl_krp	lp_npl_ck	0.003	0.004	0.749	0.454
lp_fuel	lp_npl_ck	-0.013	0.027	-0.469	0.639
lag(lp_npl_ck,1)	lp_npl_krp	0.044	0.031	1.390	0.166
lag(lp_npl_krp,1)	lp_npl_krp	0.774	0.041	19.034	0.000
constant	lp_npl_krp	-0.053	0.043	-1.244	0.215
lp_canada	lp_npl_krp	-0.004	0.014	-0.253	0.800
lp_india	lp_npl_krp	0.106	0.036	2.961	0.003
precip_npl_ck	lp_npl_krp	-0.002	0.004	-0.428	0.669
precip_npl_krp	lp_npl_krp	0.002	0.003	0.657	0.512
lp_fuel	lp_npl_krp	0.080	0.026	3.141	0.002

7.2 Causality test

Causality test is VAR based approach to explain cause-effect relationship among endogenous variables. However, the Granger-causality (Granger 1988) inference does not, of course, establish the real causation phenomena. If one of the variables is sufficiently correlated to the other so that forecast of former depends on the later to a considerable extent, then the first variable is *granger-caused* by the second one.

The Granger causality has been briefed to be useable test in certain cases of two series violating the stationarity assumption[1]. Papana et al. (2014) state that GC test can only to applied if both the series are stationary[2]. Same paper also cautioned that VAR(1) models of cointegrated endogenous series will fail to capture long-run relationships. Therefore, the authors suggest surplus lag Granger-causality test be used if the series are a nonstationary data.

Below are the results of GC test showing consequences of using both stationarity and non stationary data, with bootstrapped confidence intervals.

---

[1]https://davegiles.blogspot.com/2011/04/testing-for-granger-causality.html
[2]Mindy Mallory's blog article also suggests that series be stationary

7.3   VAR based GC test

Table 9: Granger causality test of multivariate (5 equation) VAR system of Wheat log(price) series (Precipitation trend was used as exogeneous regressor).

statistic	p.value	parameter	method
1.784	0.062	5000	Granger causality H0: lp_canada do not Granger-cause lp_india lp_npl_ck lp_npl_krp lp_fuel
2.918	0.006	5000	Granger causality H0: lp_india do not Granger-cause lp_canada lp_npl_ck lp_npl_krp lp_fuel
0.499	0.793	5000	Granger causality H0: lp_npl_ck do not Granger-cause lp_canada lp_india lp_npl_krp lp_fuel
0.580	0.629	5000	Granger causality H0: lp_npl_krp do not Granger-cause lp_canada lp_india lp_npl_ck lp_fuel
1.468	0.217	5000	Granger causality H0: lp_fuel do not Granger-cause lp_canada lp_india lp_npl_ck lp_npl_krp

8   Cointegration

8.1   Residual based

Since the food commodities are spatially linked, more of so because they occupy the same domestic market, it is obvious that factor affecting price of one inevitably affects other, especially that of same crop in a nearby market. Having evidence for nonstationarity, it is of interest to test for a common nonstationary component by means of a cointegration test (Non-stationarity is more valid for development regionwise price series).

A two step method proposed by Hylleberg et al. (1990) can be used to test for cointegration.

The procedure simply regressess one series on the other and performs a unit root test on the residuals. This test is often named after Phillips, Ouliaris, and others (1990). Specifically, `po.test()` performs a Phillips-Perron test using an auxiliary regression without a constant and linear trend and the Newey-West estimator for the required long-run variance.

The test computes the Phillips-Ouliaris test for the null hypothesis that series is not cointegrated (Trapletti and Hornik 2019).

Table 10: Phillips-Ouliaris cointegration test for Wheat log(price) series of selected district markethubs

combination	p_value	statistic
lp_canada-lp_india	0.150	-8.995
lp_canada-lp_npl_ck	0.150	-8.875
lp_canada-lp_npl_krp	0.150	-10.423
lp_canada-lp_fuel	0.150	-12.222
lp_india-lp_npl_ck	0.010	-39.607
lp_india-lp_npl_krp	0.010	-43.446
lp_india-lp_fuel	0.150	-8.948
lp_npl_ck-lp_npl_krp	0.016	-26.619
lp_npl_ck-lp_fuel	0.150	-6.692
lp_npl_krp-lp_fuel	0.150	-11.204

Note `po.test` does not handle missing values, so we fix them through imputation. It is implemented through `tidyr::fill(..., .direction = "down")`.

The test suggests that all series (Both that of wheat and rice) are cointegrated for selected pairwise combination of district markets.

The problem with this approach is that it treats both series in an asymmetric fashion, while the concept of cointegration demands that the treatment be symmetric.

The po.test() function is testing the cointegration with Phillip's Z_alpha test, which is the second residual-based test described by Phillips, Ouliaris, and others (1990). Because the po.test() will use the series at the first position to derive the residual used in the test, results would be determined by the series on the most left-hand side[3].

The Phillips-Ouliaris test implemented in the `ca.po()` function from the urca package is different. In the `ca.po()` function, there are two cointegration tests implemented, namely "Pu" and "Pz" tests. Although both the `ca.po()` function and the po.test() function are supposed to do the Phillips-Ouliaris test outcomes from both functions are completely different.

Similar to Phillip's Z_alpha test, the Pu test also is not invariant to the position of each series and therefore would give different outcomes based upon the series on the most left-hand side. On the contrary, the multivariate trace statistic of Pz test has its appeal in that the outcome won't change by the position of each series.

8.2 VAR based

The standard tests proceeding in a symmetric manner stem from Johansen's full-information maximum likelihood approach (Johansen 1991).

---

[3]https://www.r-craft.org/r-news/phillips-ouliaris-test-for-cointegration/

A general vector autoregressive model is similar to the AR(p) model except that each quantity is vector valued and matrices are used as the coefficients. The general form of the VAR(p) model, without drift, is given by:

$$\mathbf{y_t} = \; + A_1\mathbf{y_{t-1}} + ... + A_j\mathbf{y_{t-j}} + \; _t \tag{9}$$

Where is the vector-valued mean of the series, A_i are the coefficient matrices for each lag and $_t$ is a multivariate Gaussian noise term with mean zero.

At this stage we can form a Vector Error Correction Model (VECM) by differencing the series (Equation 10).

$$\Delta\mathbf{y_t} = \; + A\mathbf{y_{t-1}} + \Gamma_1\Delta\mathbf{y_{t-1}} + ... + \Gamma_j\Delta\mathbf{y_{t-j}} + \; _t \tag{10}$$

Where $\Delta\mathbf{y_t} = \mathbf{y_t} - \mathbf{y_{t-1}}$ is the differencing operator, A is the coefficient matrix for the first lag and Γ_i are the matrices for each differenced lag.

For a pth -order cointegrated vector autoregressive (VAR) model, the error correction form is (omitting deterministic components; both no intercept or trend in either cointegrating equation or test var), we may rewrite the VAR in the form of Equation 11 (Johansen 1991).

$$\Delta y_t = \Pi y_{t-1} + \sum_{j=1}^{p-1} \Gamma_j \Delta y_{t-j} + \varepsilon_t \tag{11}$$

Where,

$$\Pi = \sum_{i=1}^{p} A_i - I; \Gamma = - \sum_{j=i+1}^{p} {}_j$$

(Although, for simplicity sake, we assume absence of deterministic trends, there are five popular scenarios of including such trends in a cointegration test. All of these are described in (Johansen 1995).)

Granger's representation theorem asserts that if the coefficient matrix Π has reduced rank $r < k$, then there exist kxr matrices α and β each with rank k such that $\Pi = \alpha\beta'$ and $\beta'y_t$ is $I(0)$.

To achieve this an eigenvalue decomposition of A is carried out. The rank of the matrix A is given by r and the Johansen test sequentially tests whether this rank r is equal to zero, equal to one, through to $r = n - 1$, where n is the number of time series under test.

The null hypothesis of $r = 0$ means that there is no cointegration at all. A rank $r > 0$ implies a cointegrating relationship between two or possibly more time series.

The eigenvalue decomposition results in a set of eigenvectors. The components of the largest eigenvector admits the important property of forming the coefficients of a linear combination of time series to produce a stationary portfolio. Notice how this differs from the CADF test

Table 11: Cointegration test of log(price) series of representing national and international Wheat markets.

Cointegration relationships	10pct	5pct	1pct
r <= 4 |	10.49	12.25	16.26
r <= 3 |	22.76	25.32	30.45
r <= 2 |	39.06	42.44	48.45
r <= 1 |	59.14	62.99	70.05
r = 0 |	83.20	87.31	96.58

Table 12: ML estimation of VECM model coefficients with two cointegrating vectors, with lag order 1 and "both" (trend and constant) type deterministic regressors included in the long-term relationship.

terms	r1	r2
lp_canada	1.0000000	0.0000000
lp_india	0.0000000	1.0000000
lp_npl_ck	-2.9616586	-0.3020436
lp_npl_krp	-4.0969864	-0.6007681
lp_fuel	0.4077752	0.1070312
const	14.7780063	-0.2717388
trend	0.0338064	-0.0013949

(often known as the Engle-Granger procedure) where it is necessary to ascertain the linear combination a priori via linear regression and ordinary least squares (OLS).

In summary, the test checks for the situation of no cointegration, which occurs when the matrix $A = 0$. So, starting with the base value of r (i.e., $r = 0$), if the test statistic is greater than critical values of at the 10%, 5% and 1% levels, this would imply that we are **able** to reject the null of no cointegration. For the case r<=1, we if the calculated test statistic is below the critical values of, we are **unable** to reject the null, and the number of cointegrating vectors is between 0 and 1. The relevant tests are available in the function `urca::ca.jo()`. The basic version considers the eigenvalues of the matrix Π in the preceding equation.

Here, we employ the trace statistic – the maximum eigenvalue, or "lambdamax" test is available as well – in an equation amended by a constant term (specified by ecdet = "const"), yielding:

Results of cointegration test of the Equation 12:

Table 13: ML estimation of VECM model coefficients with two cointegrating vectors, with lag order 1, with exogeneous regressors and "trend" type deterministic regressors included in the long-term relationship.

terms	r1	r2
lp_canada	1.0000000	0.0000000
lp_india	0.0000000	1.0000000
lp_npl_ck	-1.2447300	0.0389444
lp_npl_krp	-5.0396882	-0.2652294
lp_fuel	0.5542184	0.0593170
trend	0.0284343	-0.0051257

$$
\begin{pmatrix} \Delta X_t^1 \\ \Delta X_t^2 \\ \Delta X_t^3 \\ \Delta X_t^4 \\ \Delta X_t^5 \end{pmatrix} = + \begin{pmatrix} -2.0\text{e-}03 & 0.03 \\ -5.6\text{e-}03 & -0.09 \\ -3.2\text{e-}03 & 0.09 \\ 0.01 & 0.12 \\ 9.1\text{e-}05 & 7.5\text{e-}03 \end{pmatrix} ECT_{-1} \begin{pmatrix} 0.05 \\ -0.01 \\ 0.11 \\ -0.01 \\ 7.3\text{e-}03 \end{pmatrix} + \begin{pmatrix} 0.14 & -0.05 & 0.02 & -0.11 & 0.24 \\ 0.13 & 0.02 & 0.03 & -0.02 & 0.04 \\ 0.01 & 0.02 & -0.29 & -0.01 & -0.29 \\ -0.07 & 8.9\text{e-}03 & 0.02 & -0.26 & 0.07 \\ -2.7\text{e-}03 & -0.03 & 3.3\text{e-}03 & -1.3\text{e-}03 & 0.66 \end{pmatrix} \begin{pmatrix} \Delta X_{t-1}^1 \\ \Delta X_{t-1}^2 \\ \Delta X_{t-1}^3 \\ \Delta X_{t-1}^4 \\ \Delta X_{t-1}^5 \end{pmatrix}
\tag{12}
$$

$$
\begin{pmatrix} \Delta X_t^1 \\ \Delta X_t^2 \\ \Delta X_t^3 \\ \Delta X_t^4 \\ \Delta X_t^5 \end{pmatrix} = + \begin{pmatrix} -2.0\text{e-}03(0.62) & 0.03(0.53) \\ -5.6\text{e-}03(0.04) & -0.09(6.2\text{e-}03) \\ -3.2\text{e-}03(0.26) & 0.09(0.01) \\ 0.01(2.9\text{e-}05) & 0.12(3.6\text{e-}04) \\ 9.1\text{e-}05(0.87) & 7.5\text{e-}03(0.29) \end{pmatrix} ECT_{-1} \begin{pmatrix} 0.05(0.21) \\ -0.01(0.62) \\ 0.11(6.0\text{e-}05) \\ -0.01(0.64) \\ 7.3\text{e-}03(0.18) \end{pmatrix} + \begin{pmatrix} 0.14(0.04) & -0.05(0.66) & 0.02(0.86) & -0.11(0.25) & 0.24(0.53) \\ 0.13(4.0\text{e-}03) & 0.02(0.80) & 0.03(0.56) & -0.02(0.71) & 0.04(0.88) \\ 0.01(0.79) & 0.02(0.74) & -0.29(7.2\text{e-}06) & -0.01(0.87) & -0.29(0.26) \\ -0.07(0.15) & 8.9\text{e-}03(0.90) & 0.02(0.70) & -0.26(5.5\text{e-}05) & 0.07(0.77) \\ -2.7\text{e-}03(0.77) & -0.03(0.06) & 3.3\text{e-}03(0.79) & -1.3\text{e-}03(0.92) & 0.66(2.8\text{e-}28) \end{pmatrix} \begin{pmatrix} \Delta X_{t-1}^1 \\ \Delta X_{t-1}^2 \\ \Delta X_{t-1}^3 \\ \Delta X_{t-1}^4 \\ \Delta X_{t-1}^5 \end{pmatrix}
\tag{13}
$$

$$
\begin{pmatrix} \Delta X_t^1 \\ \Delta X_t^2 \\ \Delta X_t^3 \\ \Delta X_t^4 \\ \Delta X_t^5 \end{pmatrix} = + \begin{pmatrix} -8.5\text{e-}03 & -3.9\text{e-}03 \\ -0.02 & -0.12 \\ 4.9\text{e-}03 & 0.10 \\ 0.03 & 0.13 \\ -1.7\text{e-}04 & 2.2\text{e-}03 \end{pmatrix} ECT_{-1} \begin{pmatrix} 0.07 \\ -0.02 \\ 0.09 \\ -9.9\text{e-}03 \\ 8.3\text{e-}03 \end{pmatrix} + \begin{pmatrix} 0.14 & -0.03 & 0.03 & -0.14 & 0.14 \\ 0.13 & 0.02 & 0.03 & -0.03 & -0.19 \\ 9.2\text{e-}03 & 0.02 & -0.28 & -0.01 & -0.22 \\ -0.06 & 0.01 & 0.02 & -0.25 & 0.28 \\ -4.7\text{e-}03 & -0.02 & 3.1\text{e-}03 & -3.3\text{e-}03 & 0.60 \end{pmatrix} \begin{pmatrix} \Delta X_{t-1}^1 \\ \Delta X_{t-1}^2 \\ \Delta X_{t-1}^3 \\ \Delta X_{t-1}^4 \\ \Delta X_{t-1}^5 \end{pmatrix}
\tag{14}
$$

$$
\begin{pmatrix} \Delta X_t^1 \\ \Delta X_t^2 \\ \Delta X_t^3 \\ \Delta X_t^4 \\ \Delta X_t^5 \end{pmatrix} = +
\begin{pmatrix}
-8.5\text{e-}03(0.33) & -3.9\text{e-}03(0.94) \\
-0.02(9.9\text{e-}04) & -0.12(5.8\text{e-}04) \\
4.9\text{e-}03(0.42) & 0.10(5.6\text{e-}03) \\
0.03(7.4\text{e-}07) & 0.13(1.0\text{e-}04) \\
-1.7\text{e-}04(0.88) & 2.2\text{e-}03(0.75)
\end{pmatrix} ECT_{-1}
\begin{pmatrix}
0.07(0.09) \\
-0.02(0.41) \\
0.09(2.0\text{e-}03) \\
-9.9\text{e-}03(0.72) \\
8.3\text{e-}03(0.14)
\end{pmatrix} +
\begin{pmatrix}
0.14(0.05) & -0.03(0.80) & 0.03(0.75) & -0.14(0.15) & 0.14(0.73) \\
0.13(3.9\text{e-}03) & 0.02(0.73) & 0.03(0.65) & -0.03(0.67) & -0.19(0.47) \\
9.2\text{e-}03(0.85) & 0.02(0.80) & -0.28(1.3\text{e-}05) & -0.01(0.85) & -0.22(0.44) \\
-0.06(0.15) & 0.01(0.87) & 0.02(0.76) & -0.25(1.3\text{e-}04) & 0.28(0.29) \\
-4.7\text{e-}03(0.61) & -0.02(0.08) & 3.1\text{e-}03(0.80) & -3.3\text{e-}03(0.80) & 0.60(8.8\text{e-}23)
\end{pmatrix}
\begin{pmatrix} \Delta X_{t-1}^1 \\ \Delta X_{t-1}^2 \\ \Delta X_{t-1}^3 \\ \Delta X_{t-1}^4 \\ \Delta X_{t-1}^5 \end{pmatrix}
$$

(15)

Table 14: ML estimation of VECM model coefficients with two cointegrating vectors, with lag order 1, with exogeneous regressors and with "both" (trend and constant) type deterministic regressors included in the long-term relationship.

terms	r1	r2
lp_canada	1.0000000	0.0000000
lp_india	0.0000000	1.0000000
lp_npl_ck	-1.3703295	0.0483379
lp_npl_krp	-5.5028961	-0.2318990
lp_fuel	0.6145345	0.0548494
const	12.7882176	-2.0607114
trend	0.0312277	-0.0053280

Wheat series and cointegration plots
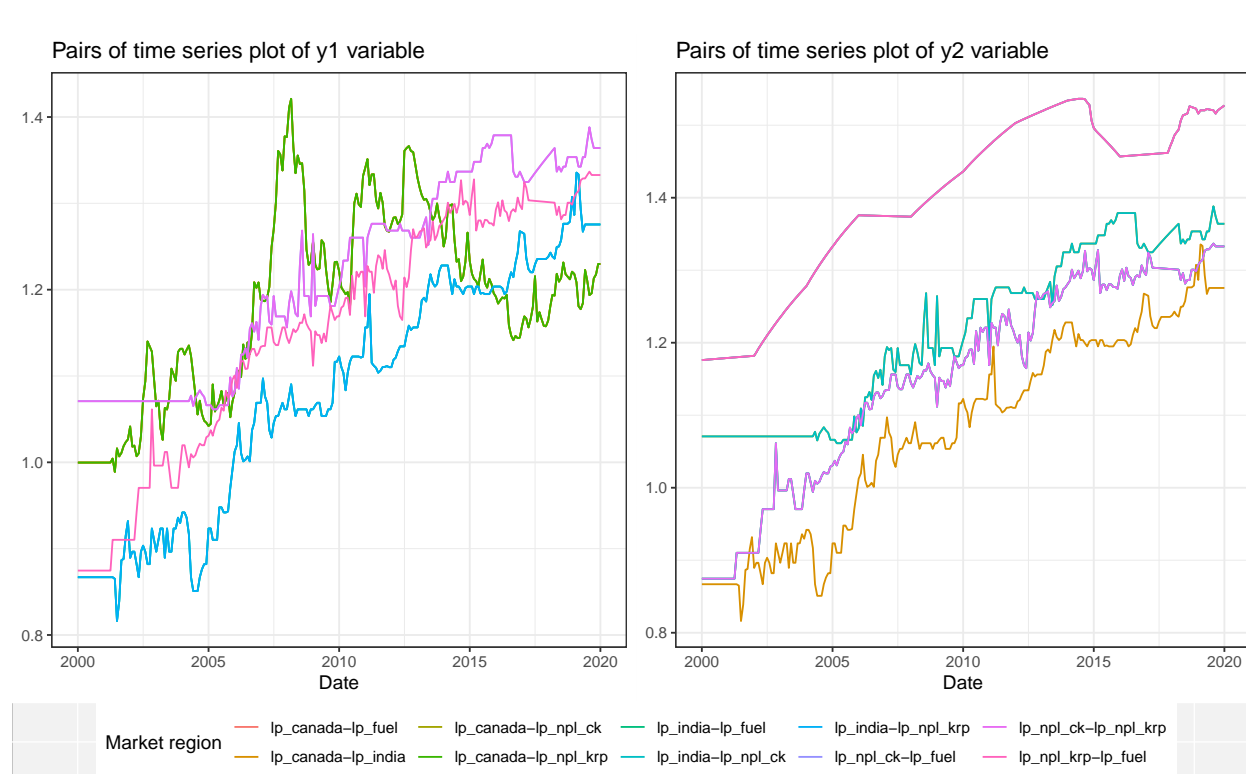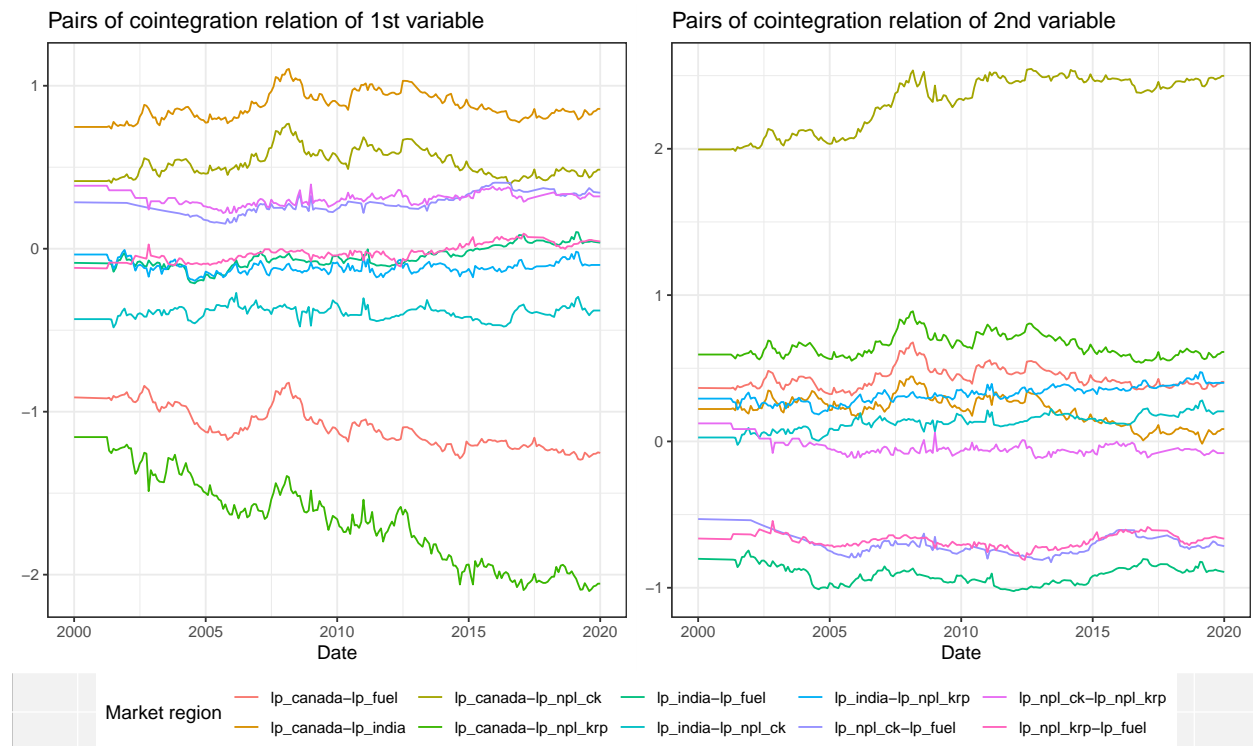
Bibliography

Granger, Clive WJ. 1988. "Causality, Cointegration, and Control." *Journal of Economic Dynamics and Control* 12 (2-3): 551–59.

Hylleberg, Svend, Robert F Engle, Clive WJ Granger, and Byung Sam Yoo. 1990. "Seasonal Integration and Cointegration." *Journal of Econometrics* 44 (1-2): 215–38.

Johansen, Søren. 1991. "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models." *Econometrica: Journal of the Econometric Society*, 1551–80.

———. 1995. "Identifying Restrictions of Linear Equations with Applications to Simultaneous Equations and Cointegration." *Journal of Econometrics* 69 (1): 111–32.

Metcalfe, Andrew V, and Paul SP Cowpertwait. 2009. *Introductory Time Series with R.* Springer.

Papana, Angeliki, Catherine Kyrtsou, Dimitris Kugiumtzis, Cees Diks, and others. 2014. "Identifying Causal Relationships in Case of Non-Stationary Time Series." *Thessaloniki: Department of Economics of the University of Macedonia.*

Phillips, Peter CB, Sam Ouliaris, and others. 1990. "Asymptotic Properties of Residual Based Tests for Cointegration." *Econometrica* 58 (1): 165–93.

Trapletti, Adrian, and Kurt Hornik. 2019. *Tseries: Time Series Analysis and Computational Finance.* https://CRAN.R-project.org/package=tseries.

Woodward, Wayne A, Henry L Gray, and Alan C Elliott. 2017. *Applied Time Series Analysis with R*. CRC press.