foreYield

Fosco Mattia Vesely

Indice

1 Introduzione

Il pacchetto R foreYield è disegnato per fornire un interfaccia utente interattiva che accompagni l'utilizzatore attraverso i vari passaggi connessi alla regressione dei risultati di simulazione per la previsione delle rese agricole[11]. La regressione richiede la disponibilità delle simulazioni effettuate per gli anni passati ed i relativi dati effettivi. Viene prevista la resa dell'anno corrente in considerazione delle simulazioni aggiornate dei regressori. Le funzioni di analisi statistica dei dati sono già disponibili in R, ma non lo è l'interfaccia e la lineare gestione dei passaggi successivi: l'impiego di foreYield agevola l'iter emancipando l'utente dagli aspetti di programmazione richiesti.

Il presente documento è destinato ad illustrare all'utente la modalità d'uso del programma. Per chi intendesse vagliare gli aspetti più tecnici si rimanda alla documentazione del pacchetto, disponibile su GitHub.

2 Installazione

Disponendo di un ambiente R già installato[1] è possibile installare foreYield con i comandi forniti in listing 1.

```
install.packages("remotes") # nel caso non sia gia' installato
library(remotes)
remotes::install_github(FoscoV/foreYield)
```

Listing 1: Procedura d'installazione entro R

Le istruzioni contenute in listing 1 provvedono automaticamente all'installazione della versione più recente di foreYield (od al suo aggiornamento) e delle relative dipendenze. Il pacchetto foreYield è basato su strumenti di analisi dati forniti in altri pacchetti. La struttura del pacchetto anzichè uno script è stata adottata, oltre che per la semplificazione nella gestione delle versioni in vista di futuri aggiornamenti, per la più comoda gestione delle dipendenze.

[1] per l'installazione si può fare riferimento al sito ufficiale del progetto CRAN

3 Note d'uso

Sono di seguito riportati i dettagli relativi alle diverse fasi dell'iter di stima. Un volta effettuata l'installazione per l'avvio della procedura è sufficiente, in una sessione di **R**, l'impiego delle istruzioni di cui in listing 2.

```
#caricamento delle librerie e dipendenze
library(foreYield)
#avvio della procedura guidata completa
virgilio()
```

Listing 2: istruzioni per l'avvio di foreYield

Una volta avviata la funzione principale secondo le indicazioni in listing 2 il resto della procedura viene affrontato interattivamente attraverso domande poste all'utente, con modalità esemplarmente riportate in figura 1.

4 Importazione dati

La procedura di importazione dei dati da un database esterno è interattivamente svolta da foreYield.
L'interfaccia richiede all'utente di indicare la posizione dei file **csv** contenenti i dati da analizzare, così come informazioni riguardanti la gestione di altri dati supportati.

Per approfondimento sulla formattazione dei dati, si faccia riferimento alla seguente sezione 4.1.

```
> virgilio()
Do you have a single file database or official and simulation split in two different files?
 1,single file
 2, double files
 (answer by numer)
1: 2
Read 1 item
Provide OFFICIAL yield database
Enter file name: eurostat.csv
Read 289 items
Provide SIMULATE yield database
Enter file name: prev.csv
Read 5684 items
OFFICIAL yield data contains information for the following CROPs:
 503 502 501 504
 Choose one: TRUE1: 504
Read 1 item
SIMULATED yield data contains information for the following CROPs:
 501 502 503 504
 Choose one: TRUE1: 504
Read 1 item
The following countries are provided in the DataBases:
 OFFICIAL: ES IT
(case sensitive)

 OFFICIAL COUNTRY:1: ES
Read 1 item
The following countries are provided in the DataBases:
 SIMULATION: ES IT1: ES
Read 1 item
It seems forecasting the year 2005 with data till the  26 th decade
Do you want to change Decade assumption?
 1: n
Read 1 item
```

Figura 1: L'iter di importazione e configurazione dei dati. Le frecce rosse indicano i punti in cui è stato digitato un input

4.1 Struttura del database

Attualmente sono supportati input in formato csv. I diversi standard a cui i csv possono fare riferimento sono autonomamente gestiti dal programma. L'unico requisito di formato è che il separatore delle cifre decimali non sia "," bensì "." onde evitare conflitti con il separatore di colonne.

I database sono file strutturati in righe e colonne. I nomi delle colonne ne indicano i contenuti. Alcune tipologie di dati richiedono uno specifico nome della colonna. Questa limitazione è resa necessaria dall'importanza di poter considerare la più ampia varietà di nomi di indicatori possibile senza imporre lunghe configurazioni dei parametri. Di seguito sono riportate le colonne il cui nome è riservato (ed obbligatorio per l'attribuzione).

YEAR contiene l'anno cui la riga viene attribuita. Data la tipologia di operazioni svolte dal programma, questa colonna è OBBLIGATORIA.

OFFICIAL_YIELD Contiene i valori reali di raccolto per l'anno in YEAR. Data la varietà di nomenclatura, sono riconosciute alcune varianti sul tema:

- official.yield
- Official yield
- Official.yield
- official_yield
- Official_yield

vengono tutte riconosciute. Questa colonna è OBBLIGATORIA

DECADE rappresenta lo stadio della stagione cui i regressori fanno riferimento. Questo campo è facoltativo. (si veda 4.2)

NUTS_CODE rappresenta la nazione o l'area di riferimento. Questo campo è facoltativo. (si veda 4.2)

CROP_NO identifica la tipologia di coltura all'interno del database; la variante STAT-_CROP_NO è riconosciuta. Questo campo è facoltativo. (si veda 4.2)

Tutti i nomi non menzionati sono disponibili per essere il nome di un regressore. Il numero di questi ultimi non è limitato in alcun modo dal programma.

Sono supportate due strutture di database, tali da incontrare le casistiche più frequentemente applicate. All'avvio viene richiesto di scegliere tra:

1. single file
2. two files

La scelta tra le due opzioni è unicamente legata alla struttura dei dati alla cui analisi di è interessati, dettagliate indicazione sulle due formattazioni sono riportate nei paragrafi 4.1.1 e 4.1.2.

Ferme restando le osservazioni sui nomi dei campi, il programma non è sensibile all'ordine.

4.1.1 Database csv in un unico file

Questa opzione importa i dati raccolti all'interno di un singolo file con una struttura del tipo riportato in tabella 1.

Tabella 1: Struttura del database **csv** in un singolo file

YEAR	DECADE	INDICATOR_CODE	INDICATOR_VALUE	OFFICIAL_YIELD
anno1	…	regressore n°1	valore	raccolto1
anno1	…	regressore n°2	valore	raccolto1
…	…	…	…	…

La struttura "long table" viene automaticamente gestita[13]. Essa, tuttavia, è basata sul nome delle colonne INDICATOR_CODE e INDICATOR_VALUE, rispettivamente per il nome del regressore ed il rispettivo valore. Questi nomi sono OBBLIGATORI per questa tipologia di file.

4.1.2 Database csv in due file

Questa opzione importa due distinti database, uno contenente i dati statistici ufficiali ed avente struttura analoga alla tabella 2 ed uno riportante i dati delle simulazioni ed analogo alla tabella 3.

Tabella 2: Struttura del database dei dati ufficiali. Solo le prime due colonne sono obbligatorie

YEAR	OFFICIAL_YIELD	NUTS_CODE	CROP_NO	colonne non considerate
anno1	raccolto1	sigla	codice	…
annno2	raccolto2	sigla	codice	…
…	…	…	…	…

Tabella 3: Struttura del database dei dati di simulazione, la colonna YEAR è obbligatoria

YEAR	DECADE	NUTS_CODE	CROP_NO	altre colonne come regressori
anno1	numero1	sigla	codice	numero1
annno2	numero1	sigla	codice	numero2
…	…	…	…	…

Il formato a due file presuppone una strutturazione "wide" della tabella, nella quale i regressori sono disposti ciascuno in una propria colonna.

4.2 Altre informazioni contenute nei database

Nel corso dell'importazione dei dati viene richiesto, nel caso siano presenti, informazioni riguardanti la coltura, la nazione e la decade di riferimento. Le richieste sono svolte al fine di isolare i dati in analisi dai dati inclusi nel database ma non rilevanti per l'analisi corrente. Per ciascuna delle interrogazioni è possibile dare risposte differenti per i dati ufficiali e per quelli di simulazione. Questa possibilità consente all'utilizzatore di impiegare diverse varianti di configurazione dei modelli distinguendole in base a codici colturali o per nazione.

NUTS_CODE e CROP_NO non sono obbligatori. Infatti database già modificati per essere limitati ad una sola coltura e/o ad un solo paese possono essere processati senza che siano dotati di queste informazioni. Data la peculiarità di siffatti database, la mancanza di questi campi viene segnalata durante la fase di importazione.

La decade, se presente, può essere selezionata, anche se di default viene suggerita la più avanzata.

L'anno per il quale viene effettuata la previsione di resa è assunto automaticamente essere il più recente, questa opzione non è configurabile. Gli anni precedenti all'ultimo sono, in ogni caso, oggetto di validazione crociata e pertanto previsti e confrontati con il risultato reale in maniera automatica.

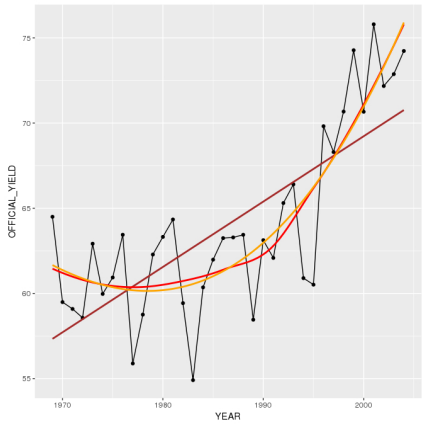


Figura 2: Un esempio dei grafici rappresentativi dell'andamento delle rese negli anni. Con regressioni (lineari e locali) per agevolarne l'interpretazione all'utente.

5 Trend

I dati ufficiali dei raccolti possono essere composti da due differenti componenti: una climatica (considerata dai modelli) ed una tecnologica (non considerata dai modelli).

L'adozione di una lunga serie di dati, seppur più suscettibile di comprendere fenomeni esogeni ai regressori, offre una maggiore validità statistica mantenendo il numero di osservazioni (gli anni considerati) nettamente maggiore del numero dei regressori.

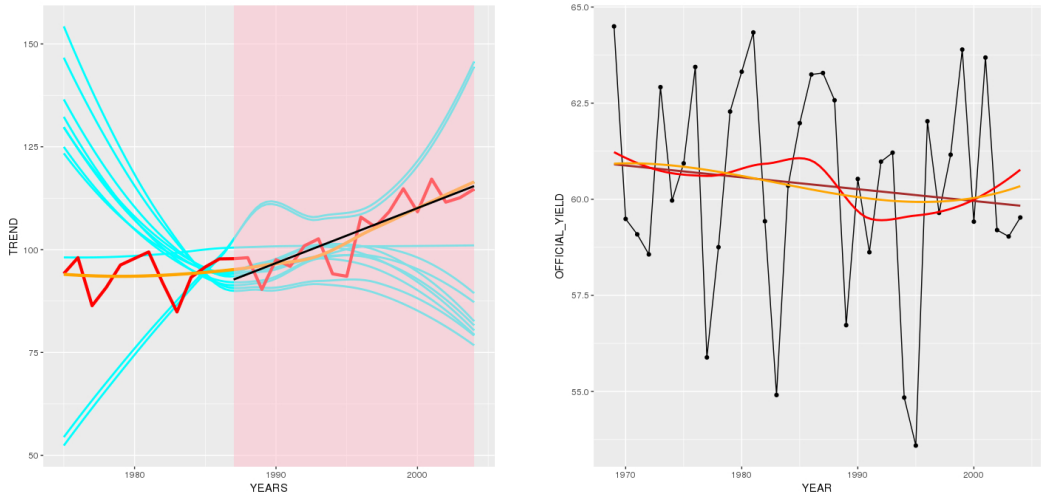
Per una corretta calibrazione del modello regressivo dei dati simulati dai modelli è necessario separare le componenti climatiche e tecnologiche. L'interfaccia fornisce una stima della probabilità che i dati ufficiali contengano un trend[9]. È importante notare che tale stima è fondata sulla sola analisi dei valori di OFFICIAL_YIELD e non discerne tra le due componenti. L'individuazione della componente tecnologica avviene sulla base della conoscenza dell'utilizzatore delle condizioni del settore agricolo in valutazione ma viene coadiuvata da una rappresentazione grafica di OFFICIAL_YIELD nel corso degli anni, analoga a quanto in figura 2.

L'utilizzatore viene interrogato circa la sua intenzione di sottrarre una componente tecnologica dai raccolti. Una risposta affermativa alla domanda avvia la procedura per l'individuazione e quantificazione della componente tecnologica. Tuttavia è lecito ritenere che il trend non sia significativo né dovuto a fattori esterni e procedere direttamente con la selezione del modello regressivo od interrompere durante le fasi della rimozione del trend. La presenza di un trend, in sé non ostacola la regressione, purchè i regressori lo possano spiegare.

5.1 Eliminazione di una componente tecnologica

Mentre le componenti climatiche sono considerate dai modelli, gli effetti tecnologici non vi sono inclusi e devono essere sottratti per garantire l'esatta regressione dei dati. Per agevolare l'individuazione della sola componente tecnologica, viene prodotto un grafico in cui sono riportati sia i raccolti sia i regressori, normalizzati ciascuno in base alla propria media (figura 3a).

Qualora si riscontri una spiegazione del trend assoluto di OFFICIAL_YIELD in base agli



(a) Confronto dei trend di resa e dei regressori.Il periodo sul cui trend si opera è evidenziato in rosa.

(b) Resa attribuibile a fattori climatici a seguito dell'elaborazione

Figura 3: Alcuni grafici di esempio delle fasi di rimozione dei trend

andamenti dei regressori, la procedura può essere interrotta[2]. In caso contrario è possibile selezionare gli anni interessati dal trend indicandone i limiti inferiore e superiore sull'asse temporale.

L'intervallo considerato viene automaticamente raffigurato sul grafico e deve essere confermato (si veda la figura 3a.

Preservare la componente climatica È possibile escludere la *sola* componente tecnologica confrontando il trend di OFFICIAL_YIELD con quello di uno o più regressori selezionabili, il cui trend medio è considerato rappresentativo della componente climatica del trend di OFFICIAL_YIELD.
L'adozione di questa soluzione consente di rimuovere la componente esogena senza ricorrere all'estrema decisione di imporre un trend nullo. All'utente è sottoposta una tabella in cui scegliere i regressori che meglio ritiene potrebbero rappresentare la componente climatica del trend.

Si osservi che la selezione di un singolo regressore porta, in seguito alla rimozione della componente tecnologica del trend, ad avere lo stesso trend lineare[3] tra il regressore ed i valori corretti di OFFICIAL_YIELD. Qualora vengano selezionati almeno due coefficienti, viene assunta come rappresentativa del trend climatico la media tra i loro andamenti. Non preoccupi particolarmente il semplice fatto che, in seguito, i coefficienti qui selezionati compaiano entro i modelli di regressione: la loro natura di regressori significativi li ha eletti secondo criteri affini.

[2] Un esempio di questa condizione potrebbe essere la condizione che si ritrova in figura 3a dove i regressori manifestano trend anche maggiori di OFFICIAL_YIELD

[3] solo entro l'arco di tempo precedentemente selezionato

Qualora gli andamenti della serie storica lo esigano, è possibile sottrarre più di una componente tecnologica da diversi intervalli temporali.

Si noti che la sottrazione di un trend tecnologico esauritosi prima della fine della serie storica sottende implicitamente che la componente tecnologica sia rimasta costantemente presente in seguito ed è pertanto automaticamente sottratta. I trend sottratti sono considerati aventi andamento lineare, in ragione della dinamica di realizzazione del trend che risulta legata alla diffusione di pratiche agricole e che pertanto si caratterizza più ragionevolmente per un andamento geometrico.

A seguito della rimozione del trend viene ripresentato il grafico di cui in figura 2 i cui dati sono però modificati, come osservabile dal confronto tra le figure 2 e 3b.

5.2 Validazione della rimozione dei trend

La rimozione dei trend tecnologici da OFFICIAL_YIELD costituisce un alterazione di valori reali necessaria per una attenta calibrazione del modello regressivo. Tali modifiche vengono validate attraverso la verifica dei due seguenti aspetti di consistenza dei dati così risultanti. Eventuali discontinuità sono segnalate nel report finale.

Endogenità dei residui Creando un modello regressivo previsionale dei valori di OFFICIAL_YIELD basato sui componenti per i quali non è segnalato un trend, vengono previsti i valori di OFFICIAL_YIELD risultanti dalla rimozione dei trend trecnologici. Viene valutata la curtosi della distribuzione degli errori rispetto ai risultati della sottrazione. Qualora fosse significativa, è interpretabile come l'indicazione che i dati impiegati per la regressione previsionale siano affetti da un agente esterno ai regressori. Quest'ultimo può essere sia un trend tecnologico non individuato sia il frutto di una infondata sottrazione di trend.

Esogenità dei trend rimossi Viene verificato che la rimozione dei trend tecnologici effettuata non aumenti lo scarto quadratico medio delle previsioni confrontando i valori di Cross Validation con le stime effettuate con lo stesso modello[4] in base ai dati per i quali non è stato indicato un trend.

6 Selezione del modello

Una volta conclusa la normalizzazione tecnologica dei valori di OFFICIAL_YIELD, questi vengono sottoposti ad analisi regressiva con diverse possibili formule di regressione lineare.[1][4]

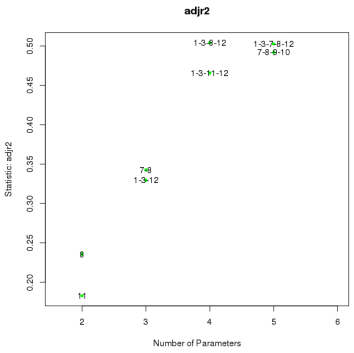
foreYield consente la creazione di regressori ottenuti dalla combinazione di due fattori regressivi. I modelli regressivi così ottenuti conseguono, solitamente, una maggiore accuratezza. Tuttavia questa pratica (denominata "enhanced" entro il programma) non è standard ed in funzione della destinazione della previsione può essere conveniente od inaccettabile.

Il numero dei possibili regressori è limitato a 4 (5, contando anche l'intercetta della retta di regressione). Per ciascun numero di regressori sono presentate le due migliori formule ed i relativi indici statistici di rappresentatività. Un esempio della tabella sottoposta

[4] La validazione è svolta in conclusione: il modello impiegato sarà quello selezionato nella sezione 6

```
  model p   rsq rss adjr2     cp     bic stderr
1         8 2 0.262 675 0.236  8.209  -2.313   4.91
2        11 2 0.211 721 0.183 10.570  -0.311   5.08
3       7-8 3 0.388 560 0.343  4.369  -4.528   4.55
4     1-3-12 3 0.376 571 0.329  4.938  -3.932   4.60
5   1-3-8-12 4 0.555 406 0.504 -1.395 -10.719   3.95
6  1-3-11-12 4 0.521 438 0.466  0.199  -8.484   4.10
7 1-3-7-8-12 5 0.571 392 0.503 -0.123  -8.397   3.96
8   7-8-9-10 5 0.562 401 0.492  0.307  -7.755   4.00

Model variables with abbreviations

                                                                                                  model
8                                                                                  RELATIVE_SOIL_MOISTURE
11                                                                                                   FSM
7-8                                                                    DEVELOPMENT_STAGE-RELATIVE_SOIL_MOISTURE
1-3-12                                                               POTENTIAL_YIELD_BIOMASS-WATER_LIM_YIELD_BIOMASS
1-3-8-12                                        POTENTIAL_YIELD_BIOMASS-WATER_LIM_YIELD_BIOMASS-RELATIVE_SOIL_MOISTURE
1-3-11-12                                                  POTENTIAL_YIELD_BIOMASS-WATER_LIM_YIELD_BIOMASS-FSM
1-3-7-8-12 POTENTIAL_YIELD_BIOMASS-WATER_LIM_YIELD_BIOMASS-DEVELOPMENT_STAGE-RELATIVE_SOIL_MOISTURE
7-8-9-10   DEVELOPMENT_STAGE-RELATIVE_SOIL_MOISTURE-TOTAL_WATER_CONSUMPTION-TOTAL_WATER_REQUIREMENT

model with largest adjr2
5

Number of observations
30

 Note: one of the accounted parameter is (Intercept)
 Select a model1:
```

(a) I modelli proposti. La freccia verde indica quale sia la chiave numerica per la selezione del modello da inserire nell'ultima riga della schermata riportata



(b) Il grafico presentato; dove sono rappresentati gli andamenti di R^2 per ciascun modello proposto nella tabella in 4a

Figura 4: Esempi estratti della procedura di selezione dei modelli regressivi

all'utente è riportato in figura 4a. Per facilitare la disponibilità di queste informazioni, esse vengono riportate in un grafico (figura 4b) automaticamente generato avente in ascisse il numero di regressori ed in ordinate l'indice $Adj R^2$. La selezione del modello avviene ad opera dell'utente sulla base delle proprie conoscenze e finalità ed è operata digitando il numero identificativo[5] corrispondente al modello selezionato.

A seguito della selezione, viene fornita indicazione della significatività dei regressori. Per ciascun regressore è presentato il corrispondente contributo al complessivo R^2[3][2].

7 Risultati

foreYield offre la stima effettuata mediante regressione delle componenti principali (PCR)[6]. Questa tecnica offre una maggiore tutela dalla collinearità della regressione. Essa non richiede alcun intervento da parte dell'utente e ne sono presentati direttamente i risultati. La PCR può costituire una buon termine di confronto per i risultati ottenuti con il modello regressivo. Affinità tra le due previsioni suggeriscono che il modello adottato non sia afflitto da fenomeni di over-fitting o collinearità che inficino la validità previsionale.

A seguito dell'indicazione del modello viene effettuata la validazione crociata[6] (Cross Validation) sia dei risultati del modello sia della Regressione per Componenti Principali (PCR). Di quest'ultima,effettuata in completa autonomia dal programma, sono presentati direttamente i risultati. I risultati della Cross Validazione sono presentati in un grafico (riportato in figura 5b) dove si hanno anche i valori aggiornati con i trend tecnologici (nel caso fossero stati rimossi).

I risultati basati sulla sola componente climatica determinante i raccolti sono:

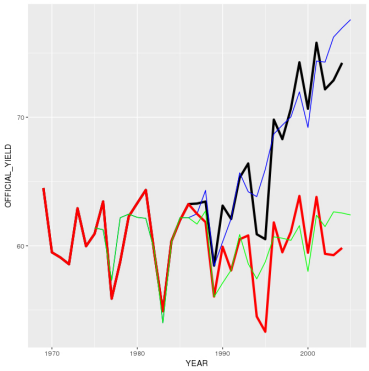
1. la previsione del modello regressivo, con relativo errore statistico

[5] evidenziato dalla freccia verde nella figura 4a

[6] La metodologia adottata per la Cross Validation è omogeneamente Leave One Out (LOO), che meglio simula le condizioni in cui i sistemi predittivi si trovano ad operare e che quindi meglio ne rappresenta la validità

```
RESPONSE

  As it is, the forecasted yield for year 2005 is 63.49 +/-
4.11 .
Confidence = 95%

 CROSS-VALIDATION
  6.54 as mean square error and  0.59 as R2.

 Principal Component Regression (PCR) predicted
  64.49 +/- 2.95  using 4 components
  64.25 +/- 2.99  using 3 components.

 Due to the marked trends, the forecasted has to be corrected with  13.18
 resulting, so, as  76.67 .



       TimeSeries statistical analysis over OFFICIAL_YIELD would bet on  73.49  +/-  6.52

 Plotted:
 GREEN: Predicted yield in Cross Validation
 RED: Data on which models are regressed
 BLUE: Predicted yield, restored trend
 BLACK: Real data
```

(a) Schermata riassuntiva degli esiti



(b) I risultati della Cross Validazione del modello, generati contestualmente alla schermata di cui in 5a

Figura 5: Esempi della presentazione dei risultati

2. il risultato della Cross Validation sui risultati di modello: errore statistico ed R^2

3. la previsione ottenuta con PCR e relativo errore in Cross Validation

Per l'applicabilità dei risultati dell'analisi è necessario considerare i trend tecnologici presenti ma non considerati nella regressione; gli effetti della componente tecnologica in OFFICIAL_YIELD, precedentemente sottratti, vengono sommati (esplicitamente).

Per offrire un termine di confronto della previsione al lordo della componente di trend tecnologico, viene presentata la previsione (aliena ad ogni simulazione agronomica) basata sulla sola serie storica dei dati di OFFICIAL_YIELD [8].

I risultati vengono forniti all'utente in una schermata riassuntiva degli esiti, esemplificativamente riportata in figura 5a.

Riferimenti bibliografici

[1] Thomas Lumley based on Fortran code by Alan Miller. *leaps: Regression Subset Selection*. R package version 3.0. 2017. URL: https://CRAN.R-project.org/package=leaps.

[2] Abraham Genizi. "Decomposition of R 2 in multiple regression with correlated regressors". In: *Statistica Sinica* (1993), pp. 407–420.

[3] Ulrike Grömping. "Relative Importance for Linear Regression in R: The Package relaimpo". In: *Journal of Statistical Software* 17.1 (2006), pp. 1–27.

[4] Richard M. Heiberger e Holland. *HH: Statistical Analysis and Data Display*. 2017. URL: https://CRAN.R-project.org/package=HH.

[5] John H. Maindonald e W. John Braun. *DAAG: Data Analysis and Graphics Data and Functions*. R package version 1.22. 2015. URL: https://CRAN.R-project.org/package=DAAG.

[6] Bjørn-Helge Mevik, Ron Wehrens e Kristian Hovde Liland. *pls: Partial Least Squares and Principal Component Regression*. R package version 2.6-0. 2016. URL: https://CRAN.R-project.org/package=pls.

[7] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2015. URL: https://www.R-project.org/.

[8] Hyndman RJ. *forecast: Forecasting functions for time series and linear models*. 2016. URL: http://github.com/robjhyndman/forecast.

[9] Adrian Trapletti e Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-37. 2017. URL: https://CRAN.R-project.org/package=tseries.

[10] W. N. Venables e B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[11] Paul Vossen, D Rijks et al. "Early crop yield assessment of the EU countries: the system implemented by the Joint Research Centre". In: (1995).

[12] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2009. URL: http://ggplot2.org.

[13] Hadley Wickham. *tidyr: Easily Tidy Data with 'spread()' and 'gather()' Functions*. R package version 0.6.1. 2017. URL: https://CRAN.R-project.org/package=tidyr.