The Art of Forecasting

Gianluca Campanella

7th June 2018

Hello!

My name is **Gianluca** [dʒanˈluːka]

What I do nowadays

I'm a Data Scientist at

Microsoft

in **Algorithms and Data Science**

What I do nowadays

I also run my own company

Estimand.com

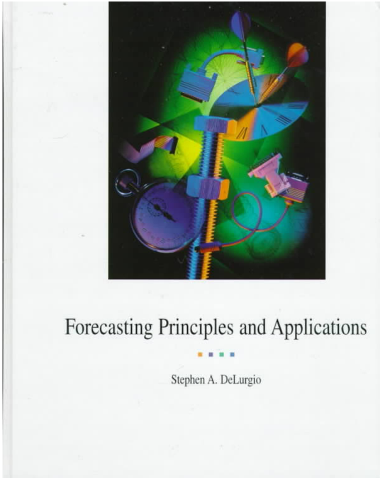
that provides

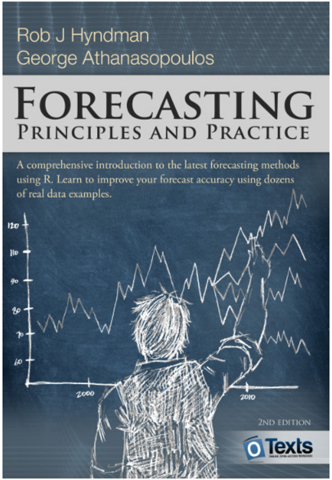
Data Science training and mentoring

Today’s slides

`https://github.com/gcampanella/ndr-2018`

References


Forecasting Principles and Applications
Stephen A. DeLurgio


Rob J Hyndman
George Athanasopoulos
FORECASTING
PRINCIPLES AND PRACTICE
A comprehensive introduction to the latest forecasting methods using R. Learn to improve your forecast accuracy using dozens of real data examples.
2ND EDITION
OTexts

Contents

What’s a time series?

Any data that change over time

- Typically continuous (including counts)
- Time gives natural ordering

What's forecasting?

Regression

- Value of $\mathbf{y}$ given values for the predictors $\mathbf{X}$
- Does not depend on time (or temporal effect is negligible)

What's forecasting?

Regression

- Value of **y** given values for the predictors **X**
- Does not depend on time (or temporal effect is negligible)

Forecasting

- Value of **y** given previous values of **y**
- Some models can also incorporate exogenous predictors

Predictability

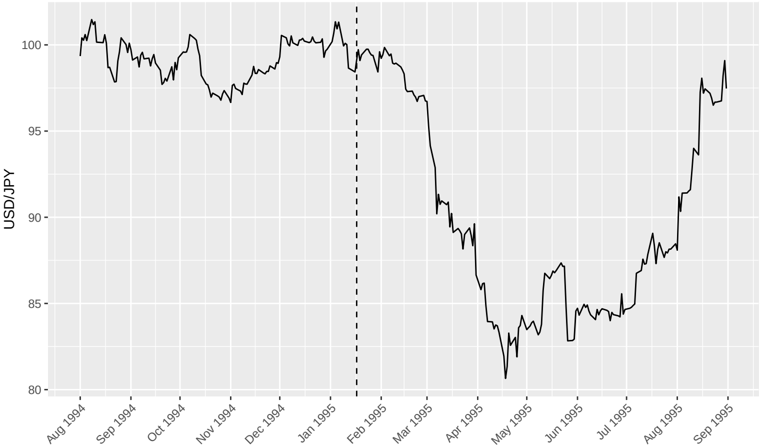
Can we forecast in changing environments?

Predictability

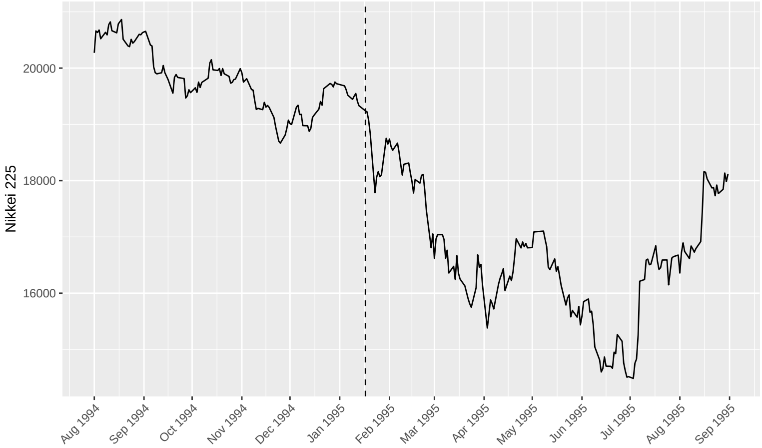
Predictability depends on…

- Availability of data
- Our understanding of contributing factors
- Whether our forecasts affect the process we're trying to forecast

A word of caution



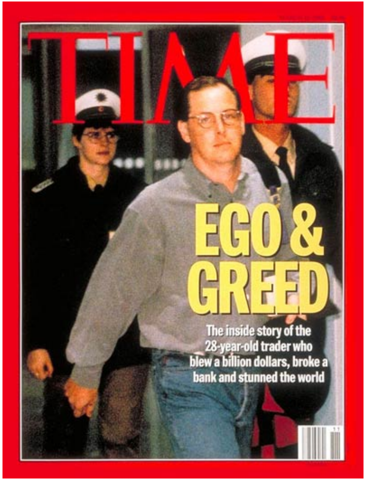
A word of caution



A word of caution

What happened?

A word of caution



Motivation

Motivation

AVVENIRE CEI NEWS SIR TV2000 RADIO INBLU FISC

Avvenire.it

SEZIONI RUBRICHE CEI PAPA OPINIONI SINODO GIOVANI

Home > Opinioni Editoriali | Il direttore risponde | Le nostre voci

Dati come di guerra nell'Italia 2015. Attenti ai morti

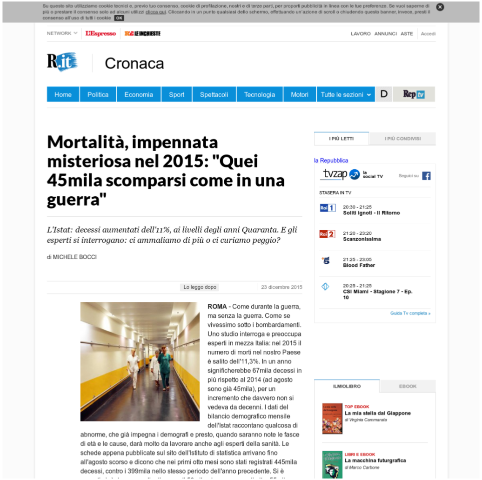
Gian Carlo Blangiardo venerdì 11 dicembre 2015

Leggendo i dati forniti dall'Istat sul totale dei morti in Italia nei primi sette mesi del 2015 – ultimo aggiornamento a tutt'oggi disponibile – si scopre un aumento di 39mila decessi rispetto agli stessi primi sette mesi del 2014. La cosa non è affatto marginale se si pensa che ciò corrisponde a un aumento dell'11% e che, se confermato su base annua, porterebbe a 664mila morti nel 2015 contro i 598mila dello scorso anno. Si tratterebbe di un aumento di ben 66mila unità, che si annuncia in gran parte concentrato sulla componente femminile (+40mila) e che verosimilmente coinvolgerà soprattutto la componente più anziana della popolazione residente nel nostro Paese.Il dato è impressionante. Ma ciò che lo rende del tutto anomalo è il fatto che per trovare un'analoga impennata della mortalità, con ordini di grandezza comparabili, si deve tornare indietro sino al 1943 e, prima ancora, occorre risalire agli anni tra il 1915 e il 1918: due periodi bellici della nostra storia che largamente spiegano dinamiche di questo tipo. Viceversa, in un'epoca come quella attuale, in condizioni di pace e con uno stato di benessere che, nonostante tutto, è da ritenersi ancora ampio e generalizzato, come si giustifica un rialzo della mortalità di queste dimensioni? È solo la naturale conseguenza del cambiamento in un popolo che diventa sempre più anziano o è (anche) un segnale di allarme rispetto a un sistema socio-sanitario che, dopo averci abituati al continuo allungamento della vita, – con guadagni sensibili anche in

OPINIONI

Secondo noi Oettinger il «re delle gaffe» che alimenta anche l'euroscetticismo

Motivation



Su questo sito utilizziamo cookie tecnici e, previo tuo consenso, cookie di profilazione, nostri e di terze parti, per proporti pubblicità in linea con le tue preferenze. Se vuoi saperne di più o prestare il consenso solo ad alcuni utilizzi clicca qui. Cliccando in un punto qualsiasi dello schermo, effettuando un'azione di scroll o chiudendo questo banner, invece, presti il consenso all'uso di tutti i cookie OK

NETWORK | L'Espresso | LE INCHIESTE | LAVORO ANNUNCI ASTE | Accedi

Cronaca

Home | Politica | Economia | Sport | Spettacoli | Tecnologia | Motori | Tutte le sezioni | D | Rep tv

Mortalità, impennata misteriosa nel 2015: "Quei 45mila scomparsi come in una guerra"

L'Istat: decessi aumentati dell'11%, ai livelli degli anni Quaranta. E gli esperti si interrogano: ci ammaliamo di più o ci curiamo peggio?

di MICHELE BOCCI

Lo leggo dopo | 23 dicembre 2015

ROMA - Come durante la guerra, ma senza la guerra. Come se vivessimo sotto i bombardamenti. Uno studio interroga e preoccupa esperti in mezza Italia: nel 2015 il numero di morti nel nostro Paese è salito dell'11,3%. In un anno significherebbe 67mila decessi in più rispetto al 2014 (ad agosto sono già 45mila), per un incremento che davvero non si vedeva da decenni. I dati del bilancio demografico mensile dell'Istat raccontano qualcosa di abnorme, che già impegna i demografi e presto, quando saranno note le fasce di età e le cause, darà molto da lavorare anche agli esperti della sanità. Le schede appena pubblicate sul sito dell'Istituto di statistica arrivano fino all'agosto scorso e dicono che nei primi otto mesi sono stati registrati 445mila decessi, contro i 399mila nello stesso periodo dell'anno precedente. Si è

I PIÙ LETTI | I PIÙ CONDIVISI

la Repubblica

tvzap la social TV — Seguici su

STASERA IN TV

20:30 - 21:25 Soliti ignoti - Il Ritorno

21:20 - 23:20 Scanzonissima

21:25 - 23:05 Blood Father

20:25 - 21:25 CSI Miami - Stagione 7 - Ep. 10

Guida Tv completa »

ILMIOLIBRO | EBOOK

TOP EBOOK La mia stella dal Giappone *di Virginia Cammarata*

LIBRI E EBOOK La macchina futurgrafica *di Marco Carbone*

Motivation

Governo, la crisi più lunga: con l'impennata dello spread si riapre l'ipotesi di Governo politico
Sale lo spread: ecco quali sono gli effetti per le imprese e l'economia reale
Spread, perché sale e perché ci deve interessare

DIETRO I DATI ISTAT

In Italia nel 2015 sono morte 54mila persone in più (+9%). Ecco le possibili cause

–di **Enrico Marro** | 25 febbraio 2016

Il rapporto Istat sugli indicatori demografici 2015 ha confermato le stime dei mesi scorsi: in Italia l'anno scorso i decessi hanno toccato quota 653mila, 54mila in più del 2014 (+9,1%). Con un tasso di mortalità, pari al 10,7 per mille, che è risultato il più alto dal secondo dopoguerra in poi. L'aumento di mortalità è concentrato tra gli anziani (75-95 anni). Come è stato possibile?

Invecchiamento popolazione e "posticipo dei decessi"
L'Istat spiega come, dal punto di vista demografico, il picco di mortalità del 2015 sia in parte dovuto a effetti strutturali connessi all'invecchiamento e in parte al posticipo delle morti non avvenute nel biennio 2013-2014. «Il picco di mortalità del 2015 porta con sé

VIDEO

30 maggio 2018
Quali effetti per i risparmi degli italiani e per le banche di casa nostra dell'impennata dello spread? Lo spiega Morya Longo

I PIÙ LETTI DI ITALIA

1. COTTARELLI AL COLLE PER «INCONTRO INFORMALE» | 29 maggio 2018
Rispunta l'ipotesi di un governo M5S-Lega. Salvini o Giorgetti in pole per Palazzo Chigi. Meloni apre
2. CRISI ISTITUZIONALE | 30 maggio 2018
Cottarelli: «C'è ipotesi di Governo politico, resto in attesa». Salvini: «voto, ma non a luglio»
3. DIETRO LA SVOLTA DI DI MAIO | 30 maggio 2018
M5S di lotta «torna» di governo: pesa la paura di perdere il voto moderato
4. NOMINE | 25 settembre 2017
Bernardo Mattarella nuovo ad della Banca del Mezzogiorno
5. LO STALLO POLITICO | 29 maggio 2018
Dall'impennata dello spread alle giravolte sul governo, cronaca di una giornata drammatica

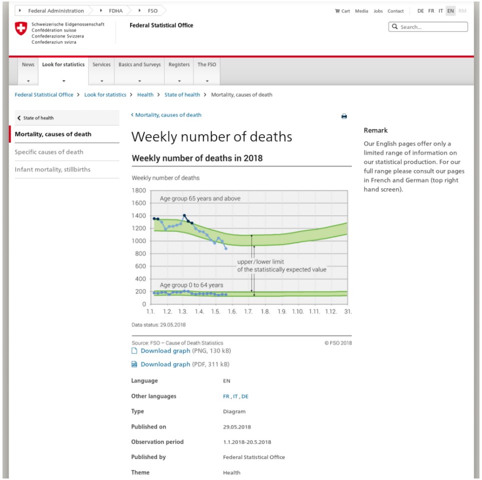
Motivation

http://demo.istat.it/

↓

https://github.com/gcampanella/istat-demographics

Motivation



Federal Administration
FDHA
FSO
Cart Media Jobs Contact
DE FR IT EN RM
Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra
Federal Statistical Office
Search...
News
Look for statistics
Services
Basics and Surveys
Registers
The FSO
Federal Statistical Office > Look for statistics > Health > State of health > Mortality, causes of death
State of health
Mortality, causes of death
Specific causes of death
Infant mortality, stillbirths
Mortality, causes of death
Weekly number of deaths
Weekly number of deaths in 2018
Weekly number of deaths
Age group 65 years and above
upper / lower limit
of the statistically expected value
Age group 0 to 64 years
Data status: 29.05.2018
Source: FSO – Cause of Death Statistics
© FSO 2018
Download graph (PNG, 130 kB)
Download graph (PDF, 311 kB)
Language EN
Other languages FR , IT , DE
Type Diagram
Published on 29.05.2018
Observation period 1.1.2018-20.5.2018
Published by Federal Statistical Office
Theme Health
Remark
Our English pages offer only a limited range of information on our statistical production. For our full range please consult our pages in French and German (top right hand screen).

Data

Original data

- Births, deaths, and net migration
- Monthly resolution from January 2004 till November 2017
- At municipality (*comune*) level
- Stratified by sex

Data

Aggregated data

- Deaths only
- Monthly resolution from January 2004 till November 2017
- At region level ($N = 20$)
- Stratified by sex

Data

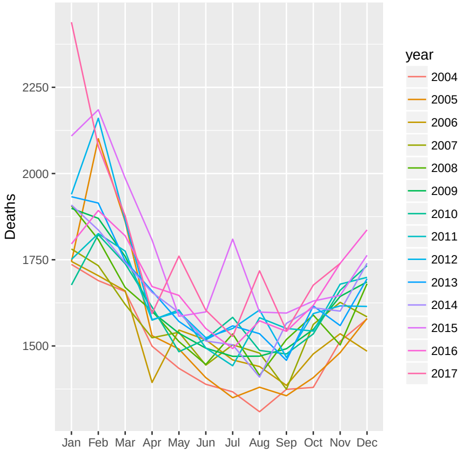
	Start	End	Length
Training	January 2004	June 2016	12.5 years
Test	July 2016	November 2017	17 months

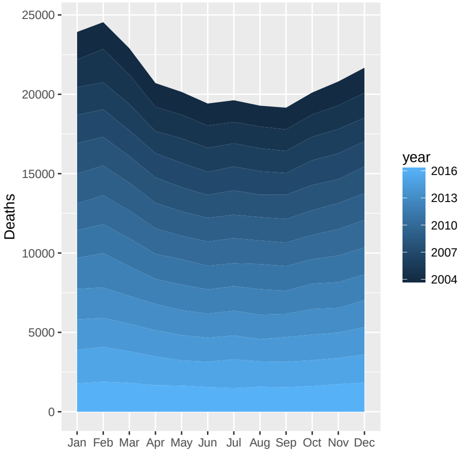
Pre-processing

Data are unnormalised monthly counts

- Boundary changes
- Population size (pre-census vs post-census)
- Calendar adjustment

Exploratory data analysis





Analysis

Family	Method	Package
Baseline	Naïve (RW)	`forecast`
	Seasonal naïve	`forecast`
	Naïve with drift	`forecast`
	Average	`forecast`
Univariate	ETS	`forecast`
	ARIMA	`forecast`
	BSTS	`bsts`
	Prophet	`prophet`
Hierarchical	HTS	`hts`

Modelling

Naïve and average methods

For all $h = 1, 2, \ldots,$

Naïve (RW)	$\hat{y}_{T+h\|T} = y_T$
Seasonal naïve with period m	$\hat{y}_{T+h\|T} = y_{T+h-m(\lfloor (h-1)/m \rfloor+1)}$
Naïve with drift	$\hat{y}_{T+h\|T} = y_T + h(y_t - y_1)/(T-1)$
Average	$\hat{y}_{T+h\|T} = \sum_{t=1}^{T} y_t / T$

Time series decomposition

Common components

- Trend-cycle T_t
- Seasonal S_t
- Remainder R_t

Additive model

$$y_t = T_t + S_t + R_t$$

Multiplicative model

$$y_t = T_t \times S_t \times R_t$$

Modelling

Exponential smoothing

Simple exponential smoothing (SES)

Given a smoothing parameter $0 \leq \alpha \leq 1$,

$$\hat{y}_{t+1|t} = \alpha y_t + (1 - \alpha)\hat{y}_{t|t-1}$$

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t && \text{(forecast)} \\ \ell_t &= \alpha y_t + (1 - \alpha)\ell_{t-1} && \text{(smoothing)} \end{aligned}$$

Holt's linear trend method

Given a smoothing parameter $0 \le \beta \le 1$,

$$\begin{aligned} \hat{y}_{t+h|t} &= \ell_t + hb_t && \text{(forecast)} \\ \ell_t &= \alpha y_t + (1-\alpha)(\ell_{t-1} + b_{t-1}) && \text{(level)} \\ b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} && \text{(trend)} \end{aligned}$$

Gardner and McKenzie's damped trend method

Given a damping parameter $0 < \phi < 1$,

$$
\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + (\phi + \phi^2 + \dots + \phi^h) b_t && \text{(forecast)} \\
\ell_t &= \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1}) && \text{(level)} \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1} && \text{(trend)}
\end{aligned}
$$

Holt–Winters' seasonal (additive) method

Given a smoothing parameter $0 \leq \gamma \leq 1$ and a frequency $m \in \mathbb{N}$,

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + hb_t + s_{t+h-m(\lfloor (h-1)/m \rfloor + 1)} && \text{(forecast)} \\
\ell_t &= \alpha(y_t - s_{t-m}) + (1-\alpha)(\ell_{t-1} + b_{t-1}) && \text{(level)} \\
b_t &= \beta(\ell_t - \ell_{t-1}) + (1-\beta)b_{t-1} && \text{(trend)} \\
s_t &= \gamma(y_t - \ell_t) + (1-\gamma)s_{t-m} && \text{(seasonality)}
\end{aligned}$$

ETS methods

- Error
    - Additive
    - Multiplicative
- Trend
    - None
    - Additive
    - Additive damped
- Seasonality
    - None
    - Additive
    - Multiplicative

$\rightsquigarrow$

$$2 \times 3 \times 3 = 18$$

possible configurations

Modelling

ARIMA models

Backshift operator $\mathcal{B}$

Let's introduce the backshift operator $\mathcal{B}$,

$$
\begin{aligned}
\mathcal{B} y_t &= y_{t-1} \\
\mathcal{B}^2 y_t &= y_{t-2} \\
&\vdots \\
\mathcal{B}^m y_t &= y_{t-m}
\end{aligned}
$$

Backshift operator $\mathcal{B}$

We can rewrite first-order differences in terms of $\mathcal{B}$,

$$\begin{aligned} y_t - y_{t-1} &= y_t - \mathcal{B}y_t \\ &= (1 - \mathcal{B})y_t \end{aligned}$$

In general, $\mathcal{B}$ follows algebraic rules,

$$\begin{aligned} (1-\mathcal{B})(1-\mathcal{B}^m)y_t &= (1 - \mathcal{B}^m - \mathcal{B} + \mathcal{B}^{m+1})y_t \\ &= y_t - y_{t-m} - y_{t-1} + y_{t-m-1} \\ &= (y_t - y_{t-m}) - (y_{t-1} - y_{(t-1)-m}) \end{aligned}$$

Autoregressive and moving average models

Autoregressive AR(p) **model of order** p

$$y_t = \beta_0 + \beta_1 y_{t-1} + \ldots + \beta_p y_{t-p} + \epsilon_t$$

Moving average MA(q) **model of order** q

$$y_t = \gamma_0 + \gamma_1 \epsilon_{t-1} + \ldots + \gamma_q \epsilon_{t-q} + \epsilon_t$$

ARIMA models

Non-seasonal ARIMA(p, d, q) **model**

$$(1 - \beta_1 \mathcal{B} - \ldots - \beta_p \mathcal{B}^p)(1 - \mathcal{B})^d y_t = \alpha + (1 + \gamma_1 \mathcal{B} + \ldots + \gamma_q \mathcal{B}^q)\epsilon_t$$

ARIMA models

Non-seasonal ARIMA(p, d, q) **model**

$$(1 - \beta_1 \mathcal{B} - \ldots - \beta_p \mathcal{B}^p)(1 - \mathcal{B})^d y_t = \alpha + (1 + \gamma_1 \mathcal{B} + \ldots + \gamma_q \mathcal{B}^q)\epsilon_t$$

Seasonal ARIMA$(p, d, q)(P, D, Q)_m$ **model**

$$(1 - \beta_1 \mathcal{B} - \ldots - \beta_p \mathcal{B}^p)(1 - \mathrm{B}_1 \mathcal{B}^m - \ldots - \mathrm{B}_P \mathcal{B}^{Pm})(1 - \mathcal{B})^d (1 - \mathcal{B}^D) y_t$$
$$= \alpha + (1 + \gamma_1 \mathcal{B} + \ldots + \gamma_q \mathcal{B}^q)(1 + \Gamma_1 \mathcal{B}^m + \ldots + \Gamma_Q \mathcal{B}^{Qm})\epsilon_t$$

Modelling

Other methods

Bayesian Structural Time Series (BSTS) models

- Introduced by S. L. Scott and H. Varian (Google)
- Ensemble method
- Structural time series model + regression component

Model evaluated

- Local linear trend
- Seasonal model with $m = 12$

Prophet

- Introduced by S. J. Taylor and B. Letham (Facebook)
- Curve fitting (similarly to GAMs)
- Decomposition into trend, seasonality, and holidays

Model evaluated

- Default settings

$\rightarrow$ No daily or weekly seasonality

Hierarchical time series models

- Introduced by R. J. Hyndman et al. (Monash University)
- Independent forecasts + aggregation at different levels
- Many different aggregation methods

Models evaluated

- Forecasting methods: ARIMA, ETS, RW
- 5 aggregation methods $\times$ 4 weighting schemes

Modelling

Measures

Scale-dependent measures

Given the prediction errors $e_{T+h} = y_{T+h} - \hat{y}_{T+h}$, …

Measure	
Mean absolute error	$\text{mean}(\lvert e_t \rvert)$
Root-mean-square error	$\sqrt{\text{mean}(e_t^2)}$

Percentage errors

Given the percentage errors $p_t = 100e_t/y_t$, ...

Measure	
Mean absolute percentage error	$\text{mean}(\lvert p_t \rvert)$
Symmetric MAPE	$\text{mean}(200\lvert y_t - \hat{y}_t \rvert/(y_t + \hat{y}_t))$

Scaled errors

Given the scaled errors...

$$q_t = \frac{e_t}{\frac{1}{T-1}\sum_{t'=2}^{T}|y_{t'} - y_{t'-1}|} \qquad \text{or} \qquad q_t = \frac{e_t}{\frac{1}{T-m}\sum_{t'=m+1}^{T}|y_{t'} - y_{t'-m}|},$$

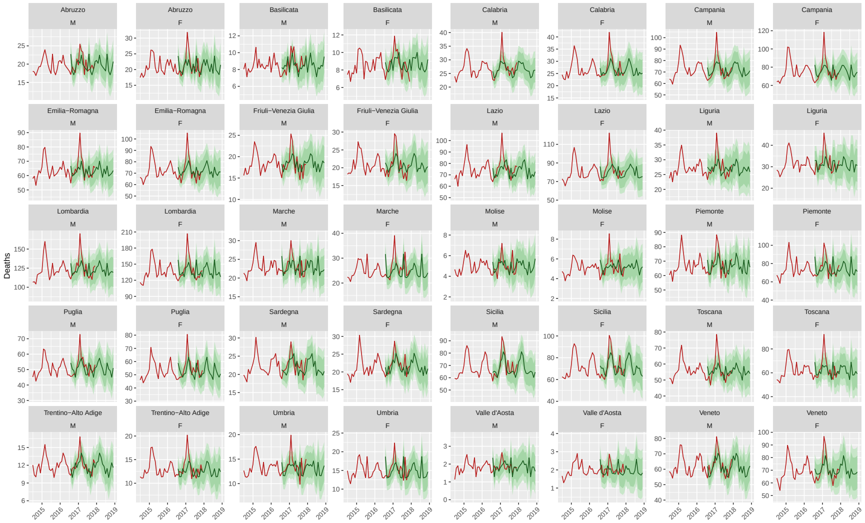
the **mean absolute scaled error** is simply mean($|q_t|$)

Interpretation

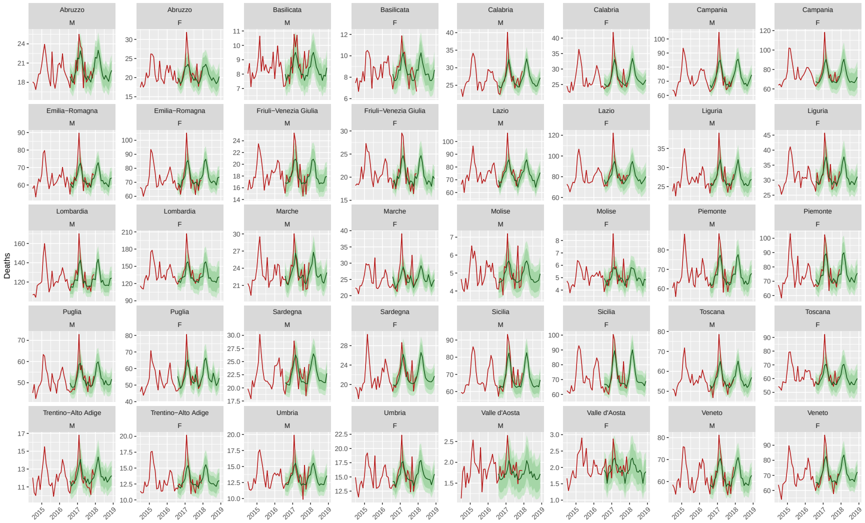
For $q_t < 1$, the forecast is better than the average (seasonal) naïve forecast (computed on the training data)

Results and recommendations

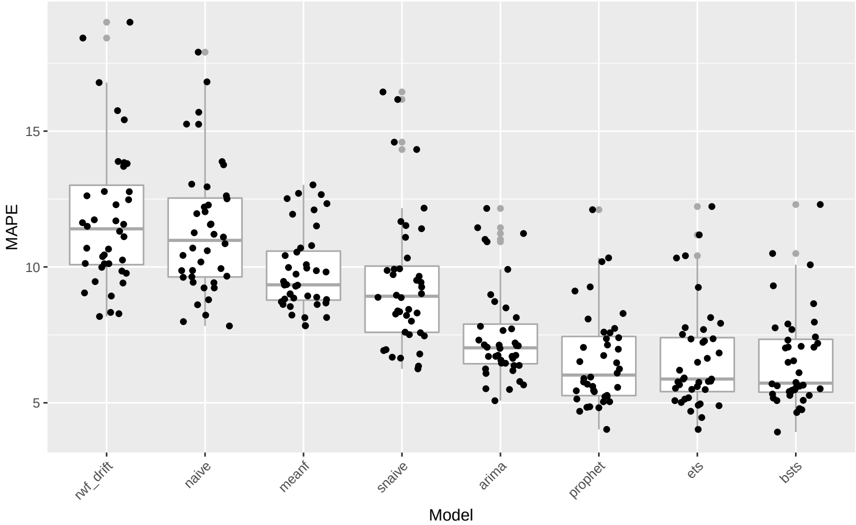
Seasonal naïve forecasts



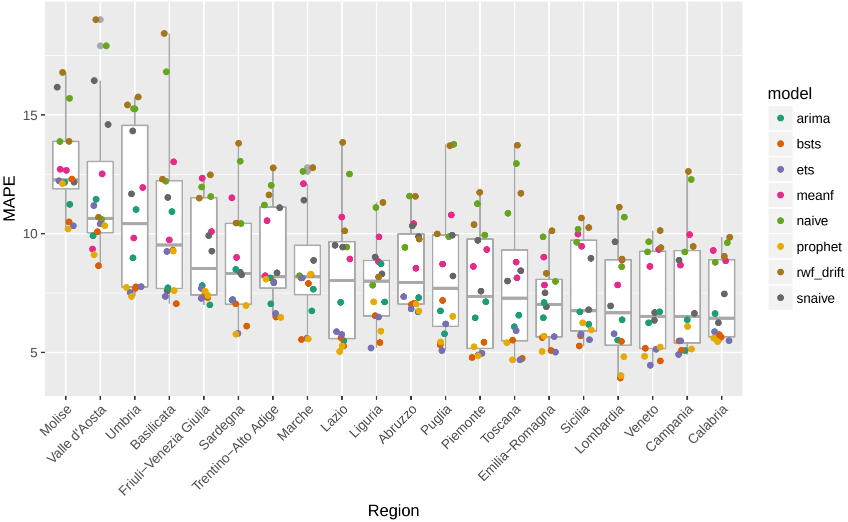
ETS forecasts



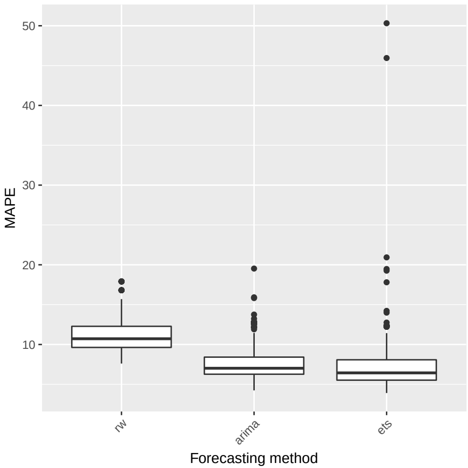
Univariate models



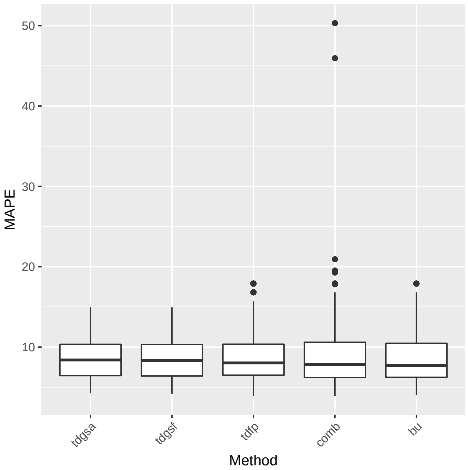
Univariate models



HTS models

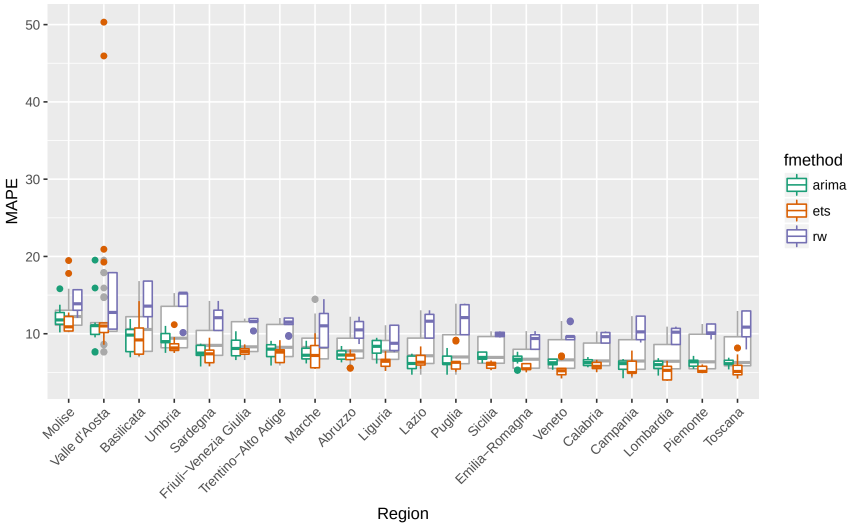


MAPE
50
40
30
20
10
rw
arima
ets
Forecasting method



MAPE
50
40
30
20
10
tdgsa
tdgsf
tdfp
comb
bu
Method

HTS models



MAPE
50
40
30
20
10
Region
Molise
Valle d'Aosta
Basilicata
Umbria
Sardegna
Friuli−Venezia Giulia
Trentino−Alto Adige
Marche
Abruzzo
Liguria
Lazio
Puglia
Sicilia
Emilia−Romagna
Veneto
Calabria
Campania
Lombardia
Piemonte
Toscana
fmethod
arima
ets
rw

And the winner is…

Method	MAPE
BSTS	6.52%
Prophet	6.58%
ETS	6.62%
HTS (bottom-up ETS)	6.62%
ARIMA	7.49%
Seasonal naïve	9.44%
Average	9.83%
Naïve (RW)	11.4%
Naïve with drift	11.8%

Lessons learned

Time series are messy!

- Temporal resolution and spacing
- Calendar adjustment
- Model evaluation and cross-validation
- Hierarchical structure

Lessons learned

Time series are fun!

- Data visualisation
- Models (often) interpretable
- Anomaly detection

Future work

- Compare even more models (including neural networks)
- Include exogenous covariates such as temperature
- Build a user interface