

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



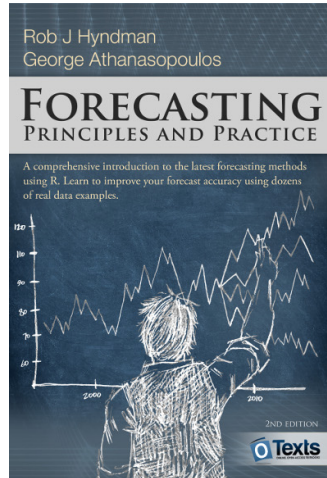
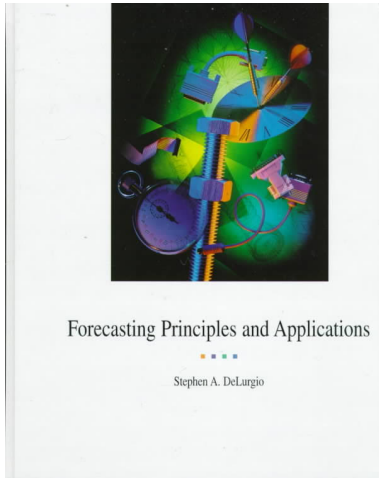
Estimand.com

that provides

Data Science training and mentoring

[https://github.com/gcampanella/
ndr-2018](https://github.com/gcampanella/ndr-2018)

References



Contents

Motivation

Modelling

Results and recommendations

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 y given values for the predictors 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

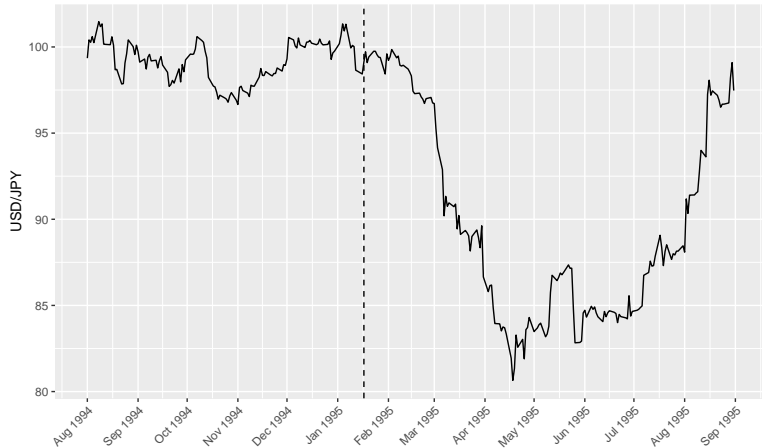
- Value of y given **previous values** of y
- Some models can also incorporate exogenous predictors

Can we forecast in changing environments?

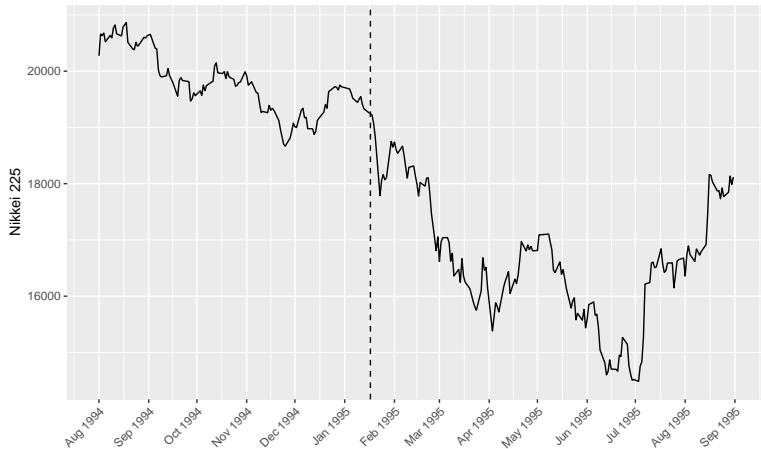
Predictability depends on...

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

A word of caution

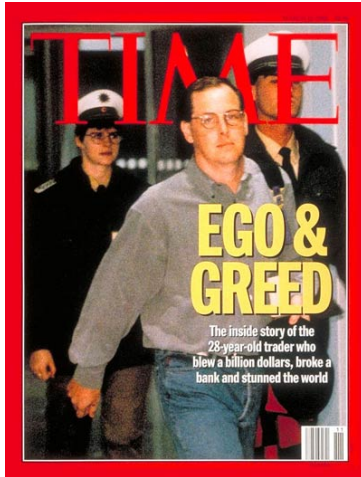


A word of caution



What happened?

A word of caution



Motivation

[AVVENIRE](#) [CEI NEWS](#) [SIR](#) [TV2000](#) [RADIO INBLU](#) [FISC](#)

Questo sito usa cookie di terze parti (anche di profilazione) e cookie tecnici. Continuando a navigare accetti i cookie. [Cambia policy](#)

accetta

[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 Bianchiardi 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'analogia 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

pubblicità

OPINIONI

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



Se il principe Filippo è il noto gaffeur della Casa reale britannica, il tedesco Guenther Gettinger è il suo «ilustre di fianco» e forse anche nell'ambito.



Looking up token.rubiconproject.com...

Se questo sito, utilizzando cookie tecnici e, previo tuo consenso, cookie di profilazione, analisi e di terze parti, per migliorare i servizi e la tua esperienza, non sei d'accordo, non puoi utilizzare questo sito. Se vuoi saperne di più o pretendere il consenso solo per alcuni utilizzi, clicca qui. Cliccando in un punto qualsiasi dello schermo, effettuando un'azione di scroll o chiudendo questo banner, invece, pretendi il consenso all'uso di tutti i cookie.

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

Rai **Cronaca**


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

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

L'Istat: decessi aumentati dell'11,3%, ai livelli degli anni Quaranta. E gli esperti si interrogano: ci ammaliamo di più o ci curiamo peggio?

di MICHELE BOCCI

Lo legge 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 attono, 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

IPU LIETI **IPU CONDIVI**

la Repubblica
tvzap **tv** **social TV** Segui su **f**

STAGIONE IN TV

1 20:30 - 21:30 **Sole igneo - Il Ritorno**

2 21:00 - 23:00 **Scanzonissima**

3 21:25 - 23:05 **Blood Father**

4 20:25 - 21:25 **CSI Miami - Stagione 7 - Ep. 19**

[Guida Tv completa >](#)

IL MQUIERO **EBDOCK**

TOP EBOOK
La mia vita da Giapponese
di Virginia Cammarata

LIBRI E EBOOK
La macchina fotografica
di Marco Carbone

Questo sito utilizza cookie, anche di terze parti, per inviarti pubblicità e servizi in linea con le tue preferenze. Se vuoi saperne di più o negare il consenso a tutti o ad alcuni cookie [clicca qui](#). Chiudendo questo banner, accettando questa pagina o cliccando su qualsiasi elemento accetti l'uso dei cookie.

OK

ATTUALITÀ PARLAMENTO POLITICA POLITICA ECONOMICA DOSSIER BLOG

Governo, le crisi più lunghe con l'imperatore dello spread si ripete l'ipotesi di Governo politico

Salvo spread: ecco quali sono gli effetti per le imprese e l'economia reale

Spread, perché sale e perché ci deve interessare

Belgio, chi è killer di guerra?

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



I 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 2016
Quali effetti per i risparmi degli italiani e per le banche di casa nostra dell'imperatore dello spread? Lo spiega Maria Longo

I PIÙ LETTI DI ITALIA

- COTTARELLI AL COLLE DEI «INCONTRO INFORMALE»** 28 maggio 2016
Risparmio e spread di un governo M5S-Lega, Salvini o Grignetti in pole per Palazzo Chigi. Milano apre
- CRISI ISTITUZIONALE** 30 maggio 2016
Cottarelli «C'è spaccato di Governo politico, resto in attesa». Salvini: no, ma non a luglio»
- DIETRO LA SVOLTA DI DI NARDI** 30 maggio 2016
M5S di lotta «vittoria» di governo: pena la paura di perdere il voto moderato
- NONE** 25 settembre 2017
Bernardo Mattarella nuovo ad della Banca del Mezzogiorno
- LO STALLO POLITICO** 28 maggio 2016
Dall'imperatore dello spread alle gioiellerie sul governo, comincia di una giornata di arruolamento

Motivation

<http://demo.istat.it/>



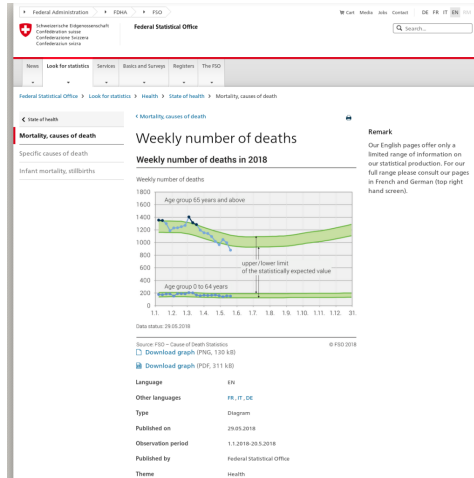
<https://github.com/gcampanella/istat-demographics>

The screenshot displays the 'demo.istat.it' website, which is a portal for Italian demographic data. The header includes the 'Geo demo istat.it' logo, the title 'Demografia in Cifre', and the Istat logo. The main content area is organized into several columns and sections:

- popolazione residente**: Includes a table for 'Popolazione Residente per età, sesso e stato civile al 1° gennaio' with data for years 2017, 2016, 2015, 2014, 2013, and 2012.
- bilancio demografico**: Includes a table for 'Bilancio Demografico a popolazione residente per sesso al 31 dicembre' with data for years 2018, 2015, 2014, 2013, 2012, and 2011.
- cittadini stranieri**: Includes a table for 'Popolazione straniera residente al 1° gennaio per età e sesso' with data for years 2017, 2016, 2015, 2014, 2013, and 2012.
- dati censuari**: Includes a section for 'I dati censuari al 9 ottobre 2011 sono disponibili alla pagina AREA PRECENSUARIA DELLA POPOLAZIONE RESIDENTE NEI COMUNI (2002-2011)'.
- L'ISTAT**: A central section providing an overview of the data available, mentioning that it includes official data on the resident population, survey results, and various indicators.
- elaborazioni**: A section for 'Tavole di Mortalità della popolazione per provincia e regione di residenza' and 'Presidenze della popolazione'.
- altri dati**: A section for 'Indicazioni e diversi' and 'La rilevazione sulle separazioni e sui divorzi'.

The website uses a clean, professional layout with a color scheme of blue, green, and orange. It provides a comprehensive overview of Italian demographic statistics and offers links to more detailed data and analyses.

Motivation



Original data

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

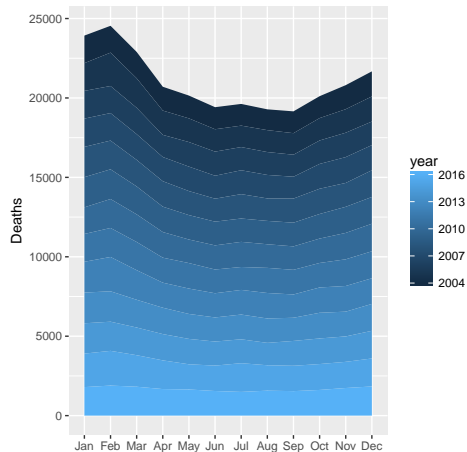
Aggregated data

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

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
	Average	forecast
Univariate	ETS	forecast
	ARIMA	forecast
	BSTS	bsts
	Prophet	prophet
Hierarchical	HTS	hts
	BYM	INLA

Modelling

Naïve and average methods

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

Method	Forecast
Naïve (RW)	Last observation $\hat{y}_{T+h T} = y_T$
Average	Average of the past $\hat{y}_{T+h T} = \frac{1}{T} \sum_t y_t$

Modelling

Exponential smoothing

Simple exponential smoothing (SES)

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

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

$$\hat{y}_{t+h|t} = \ell_t \quad \text{(forecast)}$$

$$\ell_t = \alpha y_t + (1 - \alpha) \ell_{t-1} \quad \text{(smoothing)}$$

Holt's linear trend method

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

$$\hat{y}_{t+h|t} = \ell_t + hb_t \quad \text{(forecast)}$$

$$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + b_{t-1}) \quad \text{(level)}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)b_{t-1} \quad \text{(trend)}$$

Gardner and McKenzie's damped trend method

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

$$\hat{y}_{t+h|t} = \ell_t + (\phi + \phi^2 + \dots + \phi^h)b_t \quad \text{(forecast)}$$

$$\ell_t = \alpha y_t + (1 - \alpha)(\ell_{t-1} + \phi b_{t-1}) \quad \text{(level)}$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)\phi b_{t-1} \quad \text{(trend)}$$

Holt-Winters' seasonal (additive) method

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

$$\hat{y}_{t+h|t} = \ell_t + hb_t + s_{t+h-m(\lfloor (h-1)/m \rfloor + 1)} \quad (\text{forecast})$$

$$\ell_t = \alpha(y_t - s_{t-m} + (1 - \alpha)(\ell_{t-1} + b_{t-1})) \quad (\text{level})$$

$$b_t = \beta(\ell_t - \ell_{t-1}) + (1 - \beta)b_{t-1} \quad (\text{trend})$$

$$s_t = \gamma(y_t - \ell_t) + (1 - \gamma)s_{t-m} \quad (\text{seasonality})$$

ETS methods

- **Error**

- Additive
- Multiplicative

- **Trend**

- None
- Additive
- Additive damped

- **Seasonality**

- None
- Additive
- Multiplicative

⇒

$2 \times 3 \times 3 = 18$
possible configurations

Modelling

ARIMA models

Backshift operator \mathcal{B}

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

$$\mathcal{B}y_t = y_{t-1}$$

$$\mathcal{B}^2 y_t = y_{t-2}$$

$$\vdots$$

$$\mathcal{B}^m y_t = y_{t-m}$$

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

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

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

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

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

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

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

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

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

$$\begin{aligned} (1 - \beta_1 \mathcal{B} - \dots - \beta_p \mathcal{B}^p) & (1 - \mathbf{B}_1 \mathcal{B}^m - \dots - \mathbf{B}_p \mathcal{B}^{pm})(1 - \mathcal{B})^d (1 - \mathcal{B}^D) y_t \\ & = \alpha + (1 + \gamma_1 \mathcal{B} + \dots + \gamma_q \mathcal{B}^q) (1 + \Gamma_1 \mathcal{B}^m + \dots + \Gamma_Q \mathcal{B}^{Qm}) \epsilon_t \end{aligned}$$

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
- 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

Besag–York–Mollié (BYM) model

- Introduced by J. Besag et al. (University of Washington)
- Conditional autoregressive (CAR) model
- Commonly used for disease mapping

Model evaluated

- Spatio-temporal interaction ('type IV') with first-order RWs
- Random effects model for sex

Modelling

Measures

Scale-dependent measures

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

Measure

Mean absolute error $\text{mean}(|e_t|)$

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	$\text{mean}(p_t)$
Symmetric MAPE	$\text{mean}(200 y_t - \hat{y}_t /(y_t + \hat{y}_t))$

Scaled errors

Given the **scaled** errors...

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

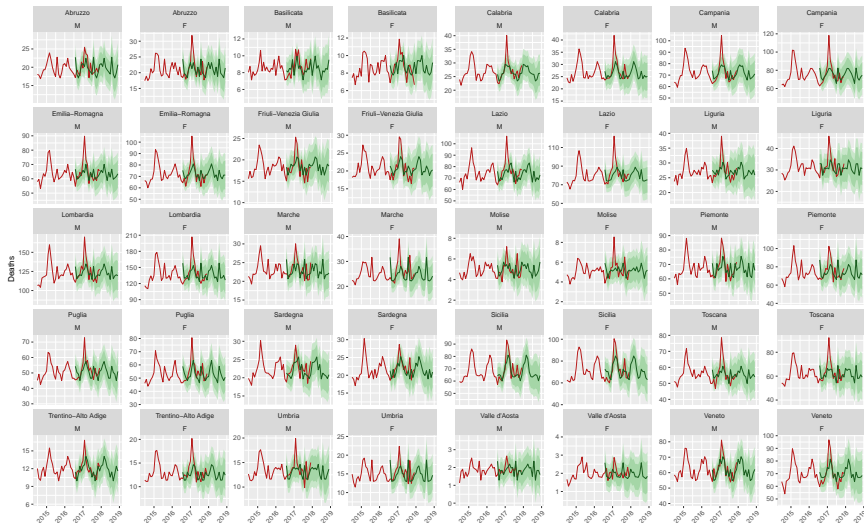
the **mean absolute scaled error** is simply $\text{mean}(|q_t|)$

Interpretation

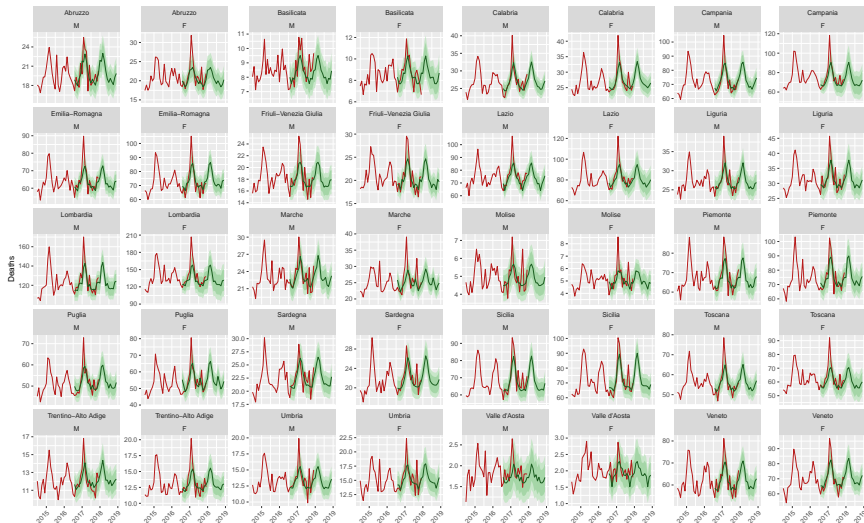
For $q_t < 1$, the forecast is better than the average (seasonal) naïve forecast (computed on the training data)

Results and recommendations

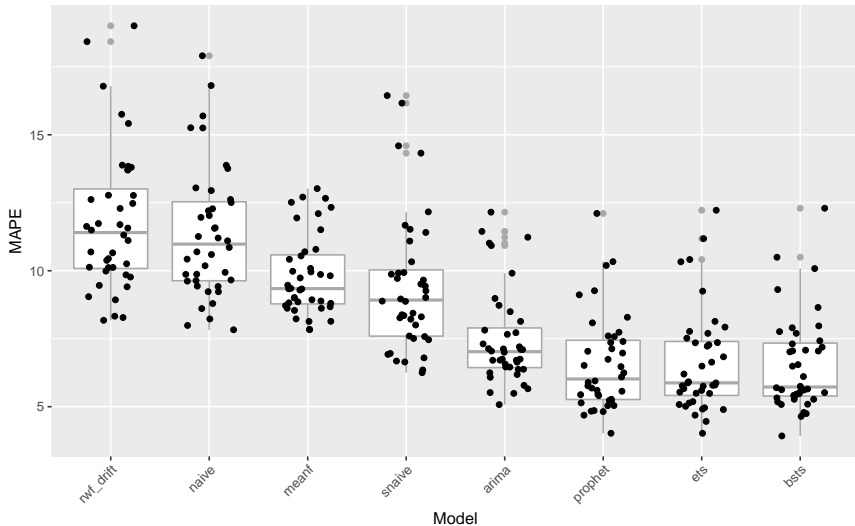
Seasonal naïve forecasts



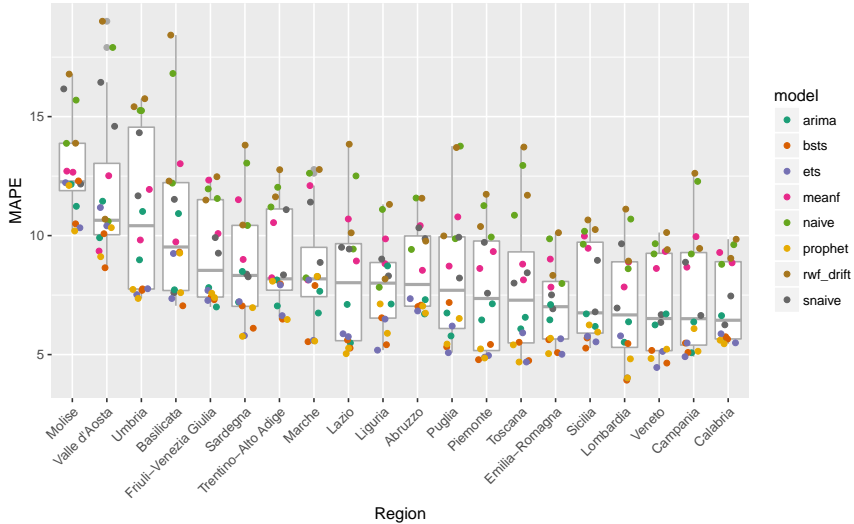
ETS forecasts



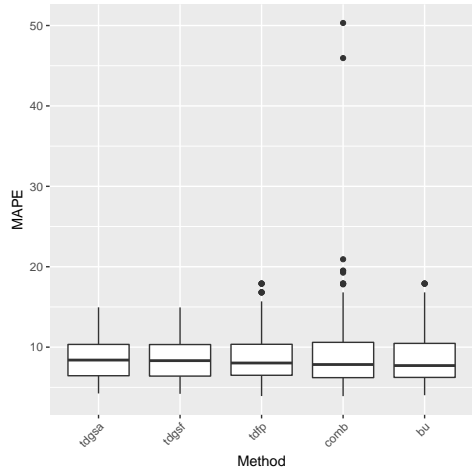
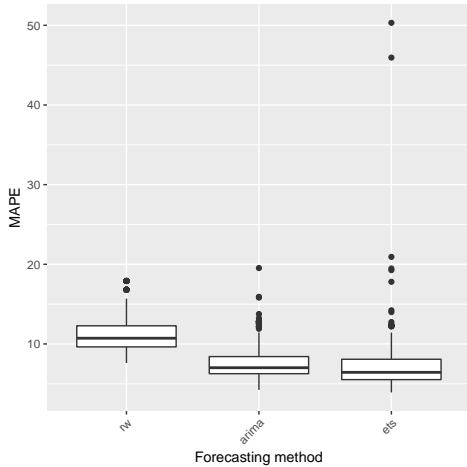
Univariate models



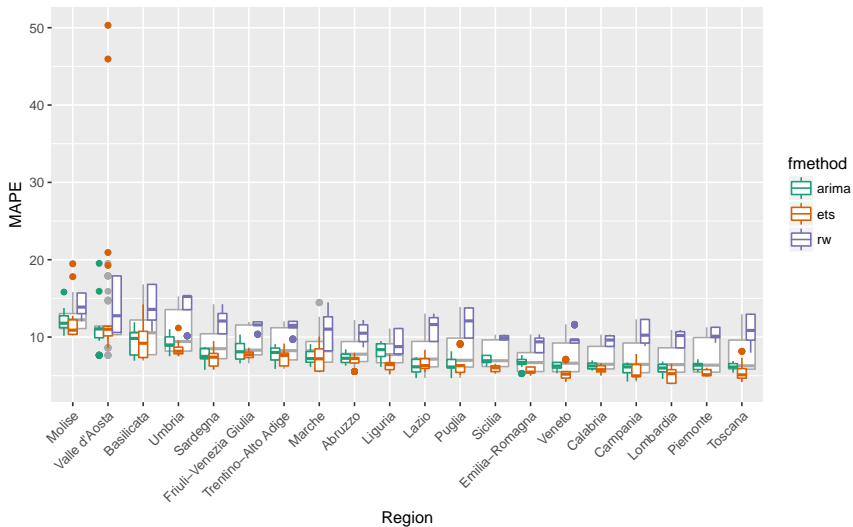
Univariate models



HTS models



HTS models



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%
RW with drift	11.8%

Lessons learned

Time series are messy!

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

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