Statistica I

Unità R: la pandemia di COVID-19 in Lombardia

Tommaso Rigon

Università Milano-Bicocca

Lezione inizialmente redatta in data

7 Novembre 2020

Unità R

Argomenti affrontati

- Pandemia di COVID-19 e infodemia (aggiornata al 7 Novembre 2020).
- I dati forniti dalla protezione civile
- Cenni alle serie storiche
- Rappresentazioni grafiche e il concetto di lisciamento
- Modelli esponenziali e previsione del contagio

La pandemia COVID-19 in Italia: una cronistoria

- **12 Gennaio 2020**. L'OMS conferma l'esistenza di un nuovo coronavirus a Wuhan, Cina. Il caso era stato portato all'attenzione dell'OMS il 31 Dicembre 2019.
- **30-31 Gennaio 2020**. Confermato il primo caso di coronavirus in Italia: si tratta di due turisti cinesi. Viene dichiarato, con delibera del Consiglio dei ministri, lo **stato di emergenza** sul territorio nazionale.
- **17 Febbraio 2020**. Un cittadino di Castiglione d'Adda (LO) si presenta all'ospedale civico di **Codogno** e viene identificato come positivo il 19 Febbraio.
- **20 Febbraio 2020**. Due persone sono state riscontrate positive per le infezioni da COVID-19 in Veneto, nel comune di **Vo'**.
- **21-22 Febbraio 2020**. Annunciato il primo Decreto legge: quarantena obbligatoria per circa 50.000 persone provenienti dai focolai attivi e isolamento di 11 comuni.

La pandemia COVID-19 in Italia: una cronistoria

- **7-9 Marzo 2020**. Un DCPM impone misure restrittive alla Lombardia ed ulteriori 14 province del Centro-Nord. Il 9 Marzo viene esteso su tutto il territorio nazionale.
- **21 Marzo 2020**. Vengono interrotte tutte le attività produttive ritenute non essenziali.
- **4 Maggio 2020**. Inizio della cosiddetta **Fase 2**: vengono allentate alcune misure di contenimento. Sono permesse le visite ai cosiddetti "congiunti".
- **15 Giugno 2020**. Ulteriore allentamento delle misure di contenimento ed inizio della **Fase 3**. Sono aperti teatri, cinema, discoteche.
- **15 Agosto 2020**. Vengono reintrodotte alcune misure restrittive. Vengono chiuse discoteche e sale da ballo.
- **13-25 Ottobre 2020**. Ulteriori misure restrittive. Viene introdotto un “coprifuoco”, i ristoranti chiudono alle 18.
- **4 Novembre 2020**. Istituzione delle zone gialle, zone arancioni e zone rosse.

I dati ufficiali italiani

- I dati relativi alla pandemia di COVID-19 sono in larga parte **pubblici**. Chiunque ha la possibilità di scaricarli e analizzarli in autonomia.

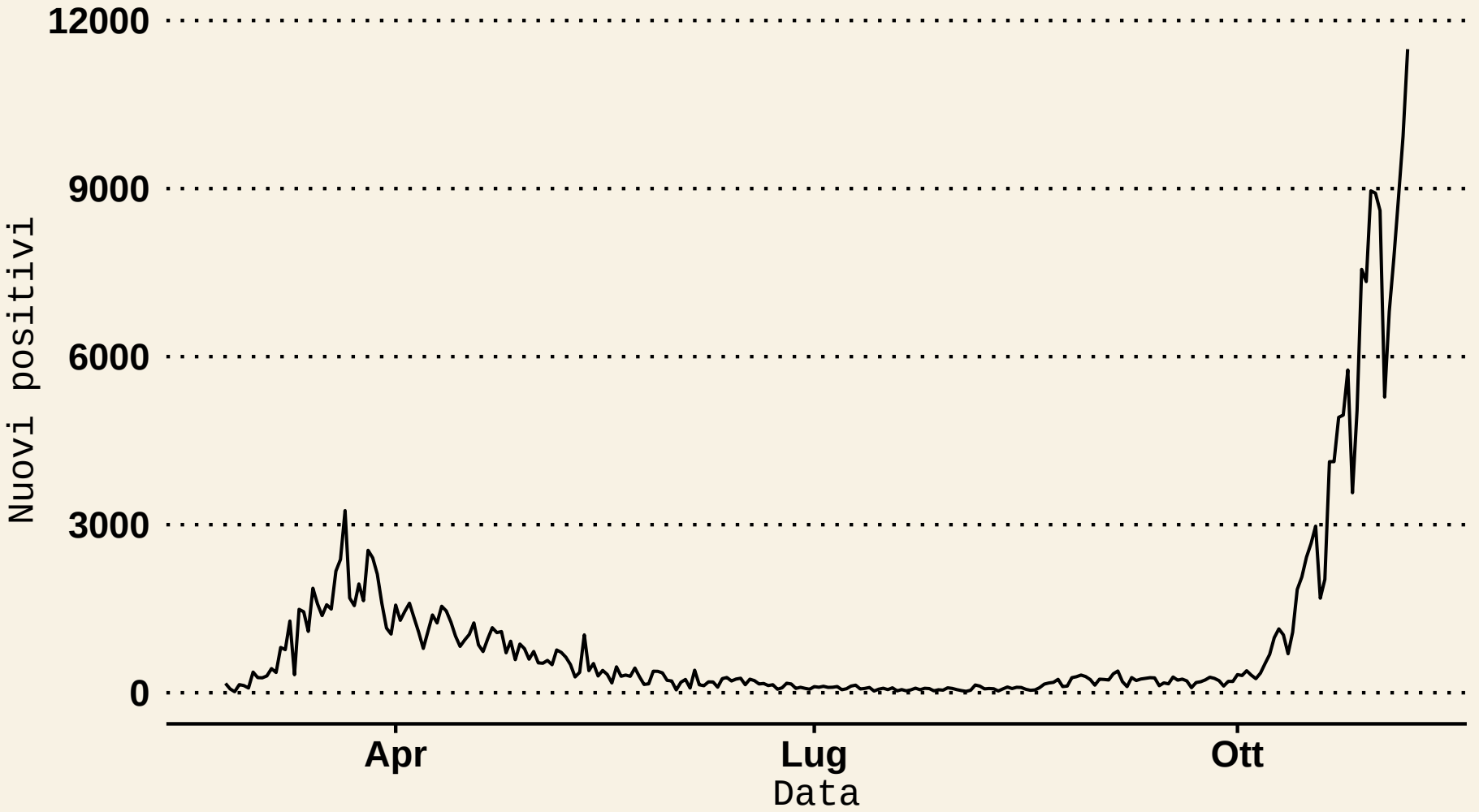
- Il **Dipartimento della Protezione Civile** aggiorna quotidianamente i dati relativi al COVID-19 nella repository: `https://github.com/pcm-dpc/COVID-19`.

- A livello nazionale, regionale e provinciale sono presenti, per tutti i giorni dall'inizio della pandemia, ad esempio:
  - Numero di contagiati giornalieri (rilevati);
  - Numero di decessi giornalieri;
  - Numero di tamponi effettuati;
  - Numeri di persone ospedalizzate;
  - Numero di persone presenti in terapia intensiva.

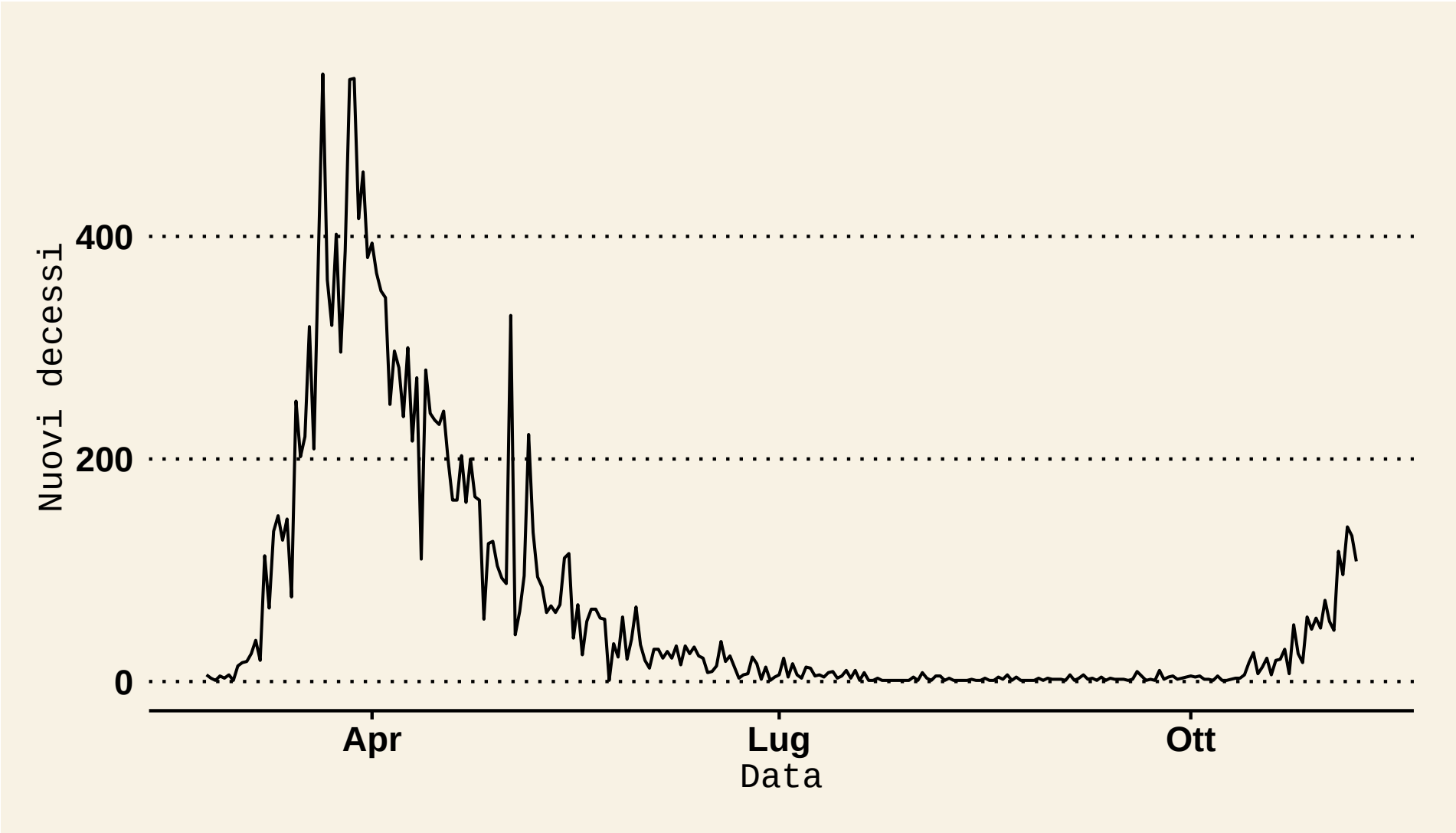
- Nelle analisi che seguono, noi faremo uso dei dati forniti dalla protezione civile e ci concentreremo solamente sulla regione **Lombardia**.

- Un'ulteriore fonte di dati pubblici (ad esempio l'età dei deceduti / contagiati) è il sito dell'**Istituto Superiore di Sanità** (ISS).

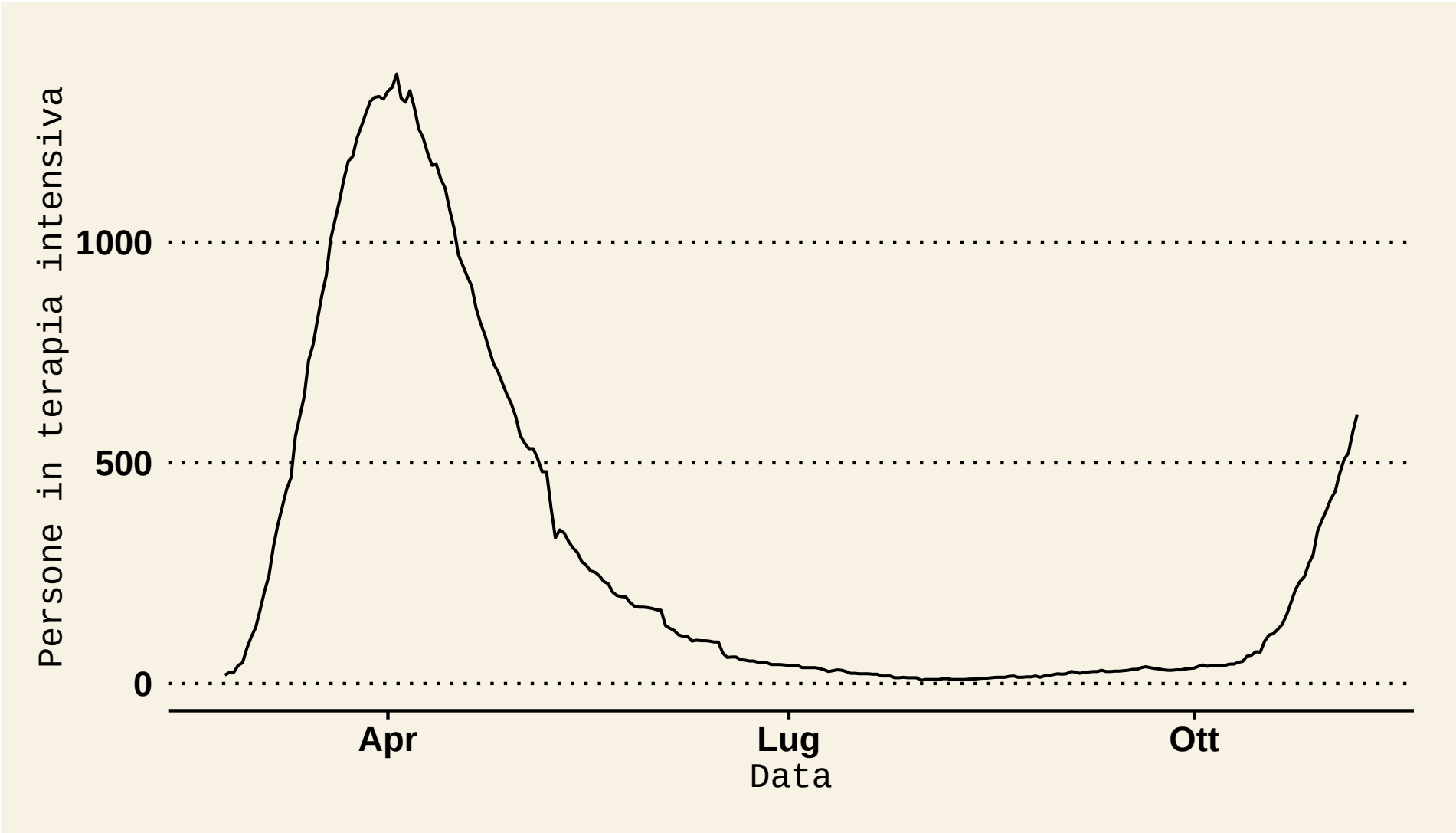
Numero di nuovi positivi (Lombardia)



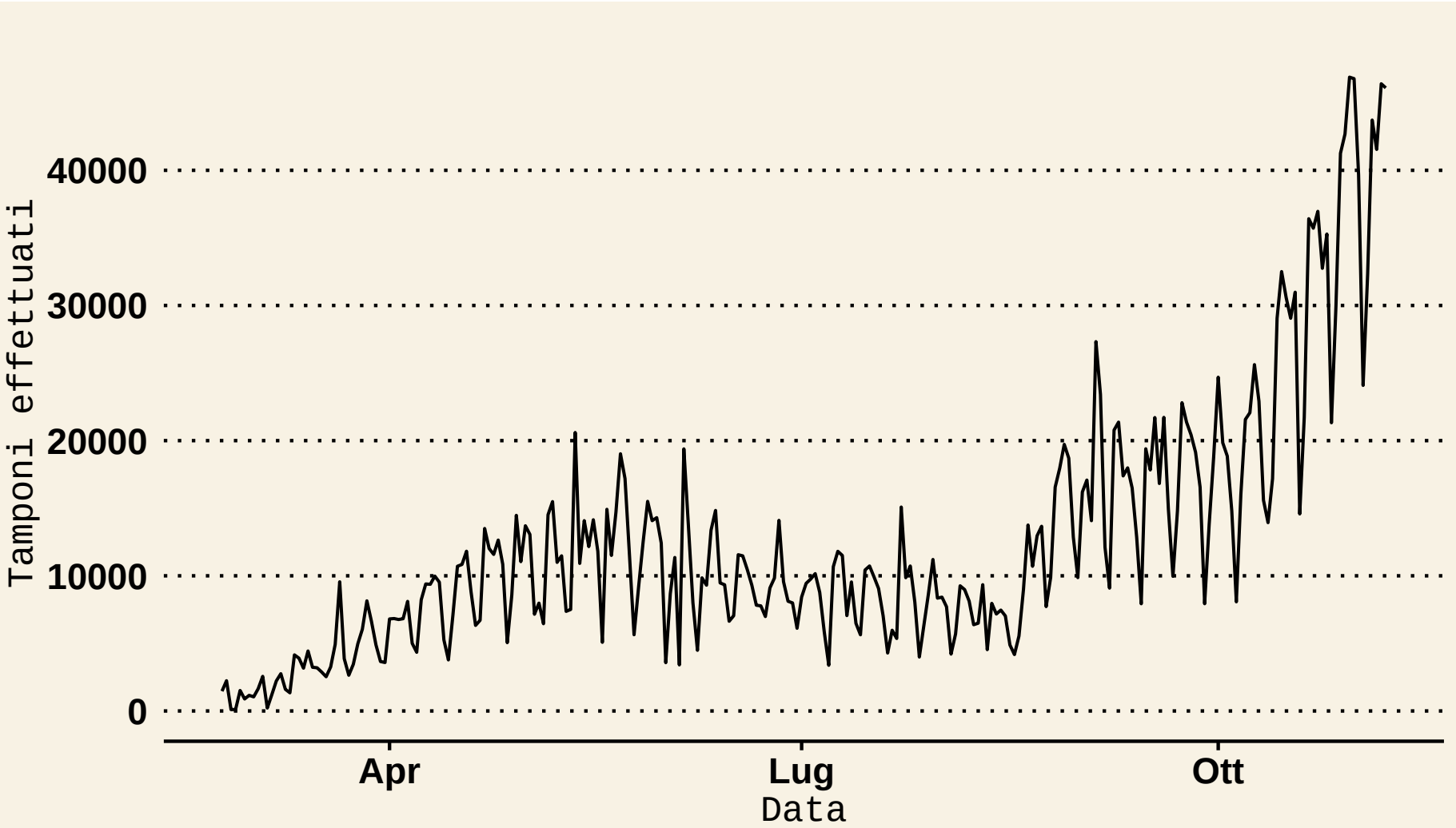
Numero di decessi (Lombardia)



Numero di pazienti in terapia intensiva (Lombardia)



Numero di tamponi effettuati (Lombardia)



Commenti ai grafici precedenti

- Forti variazioni (di contagi, decessi, etc.) avvengono in brevi periodi di tempo. Ciò è dovuto alla cosiddetta **evoluzione esponenziale** di una malattia.
- Potrebbe essere quindi utile considerare la **scala logaritmica**.
- I dati manifestano delle **fluttuazioni casuali** che complicano l'interpretazione del messaggio. Non ha quindi **nessun senso** analizzare le **variazioni giornaliere**.
- Oltretutto, ad esempio al Lunedì il numero di nuovi contagiati è sistematicamente inferiore rispetto al trend.
- **Attenzione** all'interpretazione dei dati. I nuovi positivi di Ottobre non sono confrontabili con quelli di Marzo a causa del minor numero di tamponi effettuati.

Problematiche dei dati ufficiali

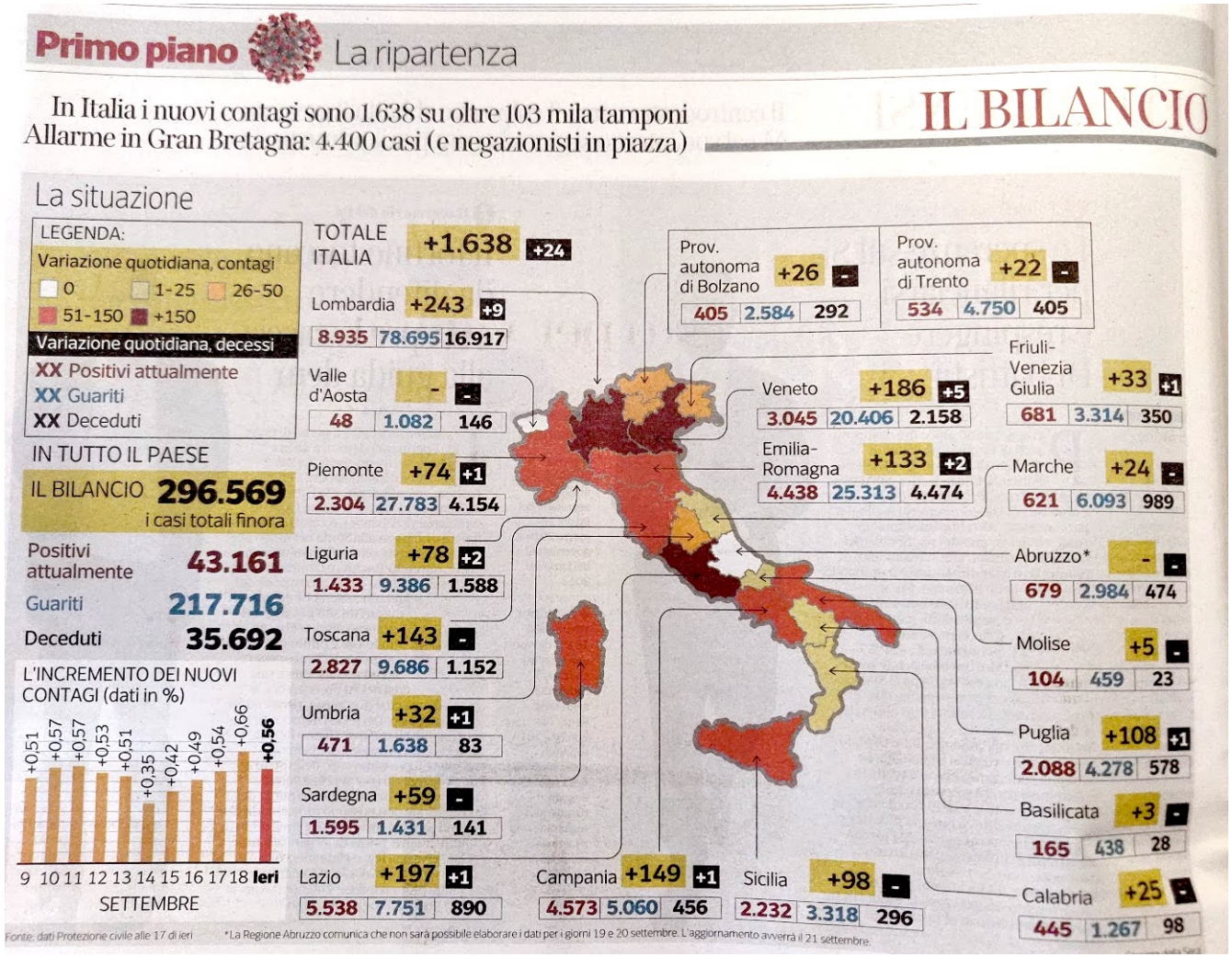
- I tamponi sono uno strumento **medico** e **non statistico**. Il tampone viene effettuato in base alla necessità, con minore attenzione alla rappresentatività del campione.
- Inoltre, i tamponi sono disponibili in quantità finite, per cui vanno necessariamente gestiti con parsimonia.
- Infine, l'ampia frazione di **contagiati asintomatici** complica la gestione del monitoraggio.
- Di conseguenza, ci sono delle oggettive **difficoltà nel monitoraggio** se questo è basato unicamente sui tamponi e sul numero dei contagiati.
- Altre complicazioni: errori di trascrizione, ritardo nelle comunicazioni da parte delle regioni, etc.

Infodemia e comunicazione

- I dati sono pertanto di **difficile interpretazione**.
- Questo tuttavia non implica che sia impossibile estrarne delle informazioni.
- Nonostante il messaggio contenuto nei dati sia spesso abbastanza **univoco**, in alcuni canali di informazione (giornali, telegiornali, social network) questi dati sono stati usati per sostenere le opinioni più diverse, anche opposte tra loro.
- Lo sfruttamento delle emozioni e la scarsa attinenza ai fatti (ovvero ai dati) ha quindi generato una vera e propria **infodemia**.

Dati, giornalismo e social network

Dati, giornalismo e social network



Primo piano La ripartenza

In Italia i nuovi contagi sono 1.638 su oltre 103 mila tamponi
Allarme in Gran Bretagna: 4.400 casi (e negazionisti in piazza)

IL BILANCIO

La situazione

LEGENDA:
Variazione quotidiana, contagi
0 | 1-25 | 26-50 | 51-150 | +150
Variazione quotidiana, decessi
XX Positivi attualmente
XX Guariti
XX Deceduti

IN TUTTO IL PAESE

IL BILANCIO 296.569 i casi totali finora

Positivi attualmente 43.161
Guariti 217.716
Deceduti 35.692

L'INCREMENTO DEI NUOVI CONTAGI (dati in %)

+0,51 +0,57 +0,57 +0,53 +0,51 +0,35 +0,42 +0,49 +0,54 +0,66 +0,56

9 10 11 12 13 14 15 16 17 18 Ieri
SETTEMBRE

TOTALE ITALIA +1.638 +24

Lombardia +243 +9 — 8.935 | 78.695 | 16.917
Valle d'Aosta - - — 48 | 1.082 | 146
Piemonte +74 +1 — 2.304 | 27.783 | 4.154
Liguria +78 +2 — 1.433 | 9.386 | 1.588
Toscana +143 - — 2.827 | 9.686 | 1.152
Umbria +32 +1 — 471 | 1.638 | 83
Sardegna +59 - — 1.595 | 1.431 | 141
Lazio +197 +1 — 5.538 | 7.751 | 890
Campania +149 +1 — 4.573 | 5.060 | 456
Sicilia +98 - — 2.232 | 3.318 | 296

Prov. autonoma di Bolzano +26 - — 405 | 2.584 | 292
Prov. autonoma di Trento +22 - — 534 | 4.750 | 405
Veneto +186 +5 — 3.045 | 20.406 | 2.158
Friuli-Venezia Giulia +33 +1 — 681 | 3.314 | 350
Emilia-Romagna +133 +2 — 4.438 | 25.313 | 4.474
Marche +24 - — 621 | 6.093 | 989
Abruzzo* - - — 679 | 2.984 | 474
Molise +5 - — 104 | 459 | 23
Puglia +108 +1 — 2.088 | 4.278 | 578
Basilicata +3 - — 165 | 438 | 28
Calabria +25 - — 445 | 1.267 | 98

Fonte: dati Protezione civile alle 17 di ieri
*La Regione Abruzzo comunica che non sarà possibile elaborare i dati per i giorni 19 e 20 settembre. L'aggiornamento avverrà il 21 settembre.

Dati, giornalismo e social network

Il Sole 24 ORE
3 h ·

Superati ancora i numeri del 21 marzo, picco più alto della prima ondata, con 6.557 infezioni su 26.336 test

ILSOLE24ORE.COM
Coronavirus, ultime notizie. In Italia nuovo record di contagi e tamponi: +8.804 positivi su 162....

Il ruolo dello statistico, il ruolo dello scienziato

- L'analisi dei dati COVID-19 è difficile e complessa. Quindi ogni considerazione richiede
  - estrema cautela
  - un'attenta riflessione sulle ipotesi sottostanti.
- Lo statistico tuttavia non deve sottrarsi alla sfida. La scienza (dei dati) non è infatti una collezione di risposte, bensì **metodo**.
- Alle **scetticismi antiscientifici** dei movimenti "no-mask" e "no-vax", è indubbiamente preferibile un metodo, seppur imperfetto, che sia basato sui **fatti** e il **raziocinio**.
- *"E' semplice mentire con la statistica, ma è molto più semplice mentire senza di essa"*. **Frederick Mosteller**.

Le serie storiche

- Le **serie storiche** sono dei dati indicizzati dal **tempo**, come nel caso dei contagiati, decessi, etc.
- Il tempo (giorni, ore, minuti, etc.) viene codificato con dei numeri $t = 1, 2, \ldots, T$.
- In questo contesto, al posto dei dati $x_1, \ldots, x_n$ useremo la notazione $x_1, \ldots, x_T$.

Variabile x	x_1	$\ldots$	x_t	$\ldots$	x_T
Tempo	1	$\ldots$	t	$\ldots$	T

- Ad esempio, avremo che

Variabile `decessi`	6	$\ldots$	12	$\ldots$	108
Tempo	1	$\ldots$	100	$\ldots$	258
Data	24-02-2020	$\ldots$	02-06-2020	$\ldots$	7-11-2020

Dati giornalieri e rumore

- Per le più varie ragioni (errori di tracciamento, naturale variabilità del processo di contagio), i dati grezzi sono **rumorosi**.
- Il primo importante obiettivo che ci poniamo è quindi identificare un **trend**.
- È del tutto inutile commentare le variazioni giornaliere. Viceversa, l'analisi del trend permette di capire la **direzione** della pandemia.
- Ad esempio, consente di capire se il contagio si sta espandendo, se sta recedendo o se ha raggiunto il suo picco.
- Esistono svariati metodi di identificazione del trend, qui presentiamo uno dei più semplici, basato sul concetto di media aritmetiche locali.

Filtri lineari e medie mobili

- Una serie storica $y_1, \ldots, y_T$ **filtrata** si ottiene come

$$y_t = \sum_{i=-h}^{k} \tilde{w}_i x_{t+i}, \qquad t = 1, \ldots, T,$$

dove $x_1, \ldots, x_T$ è la serie storica originale e $\tilde{w}_i \geq 0$ sono dei pesi normalizzati, ovvero tali che

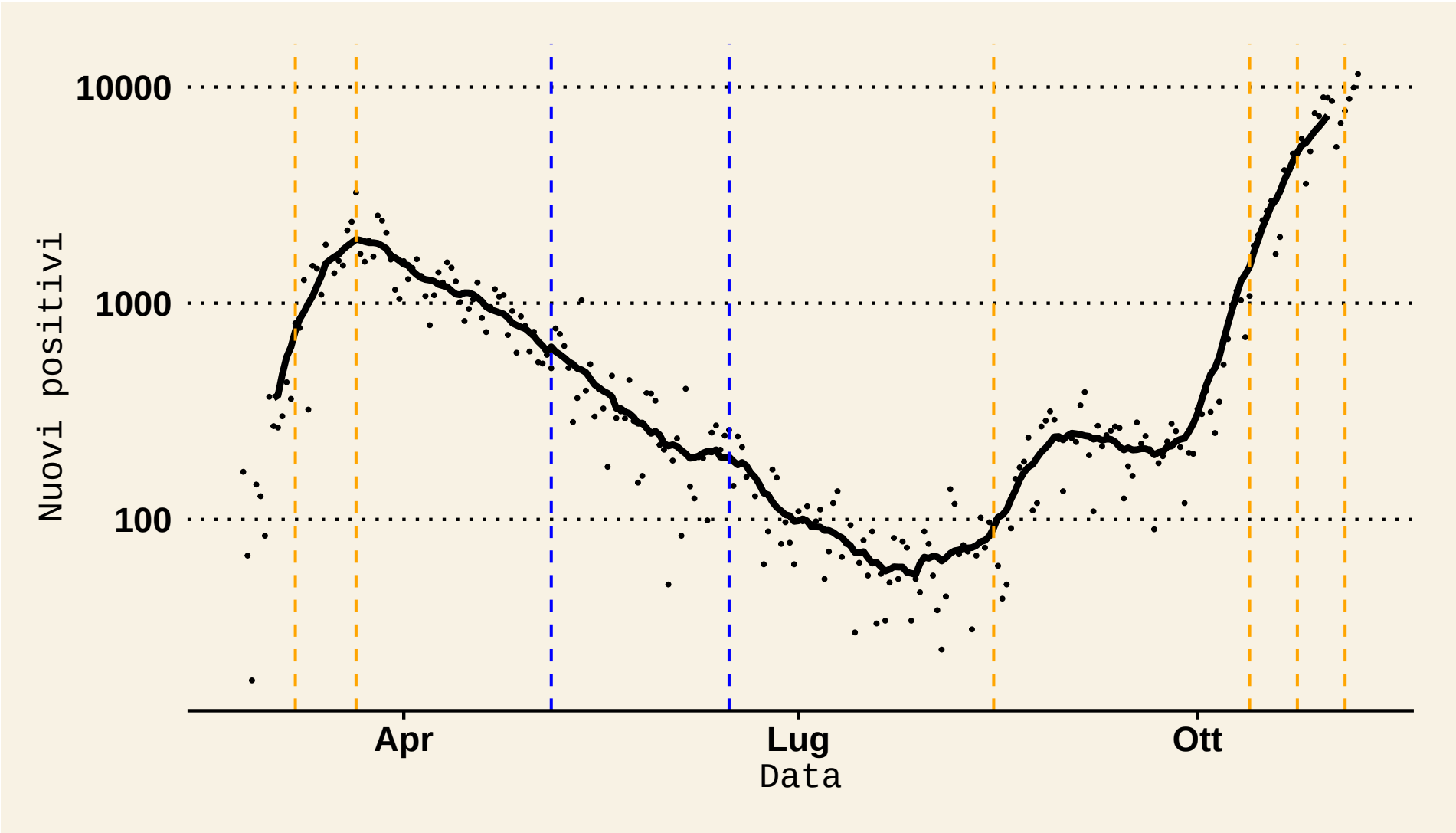
$$\sum_{i=-h}^{k} \tilde{w}_i = 1.$$

- Tale operazione viene detta **filtro lineare** oppure **media mobile**.

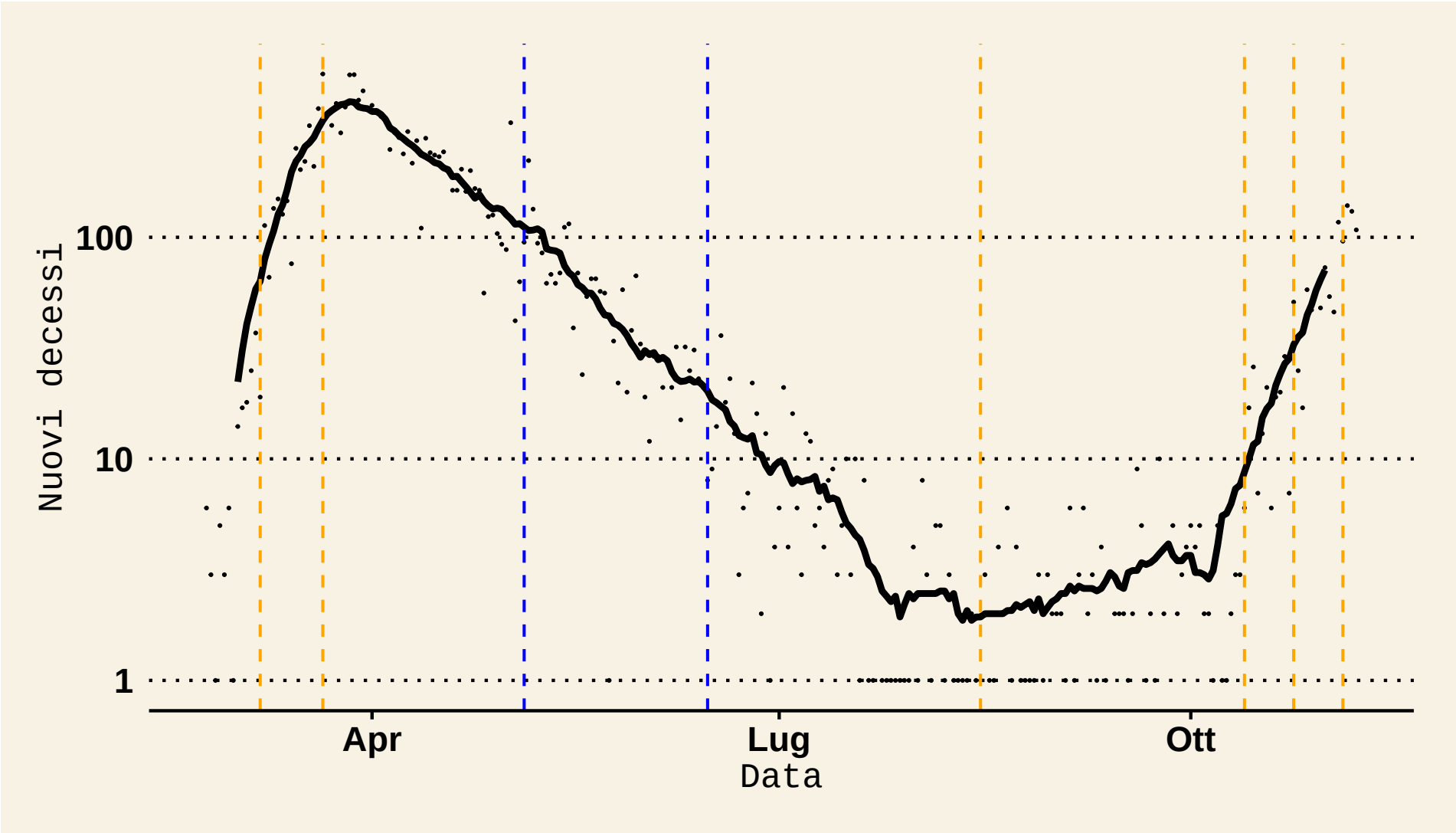
- Nel nostro contesto, utilizzeremo il seguente filtro:

$$y_t = \frac{1}{15} \sum_{i=-7}^{7} x_{t+i}, \qquad t = 1, \ldots, T.$$

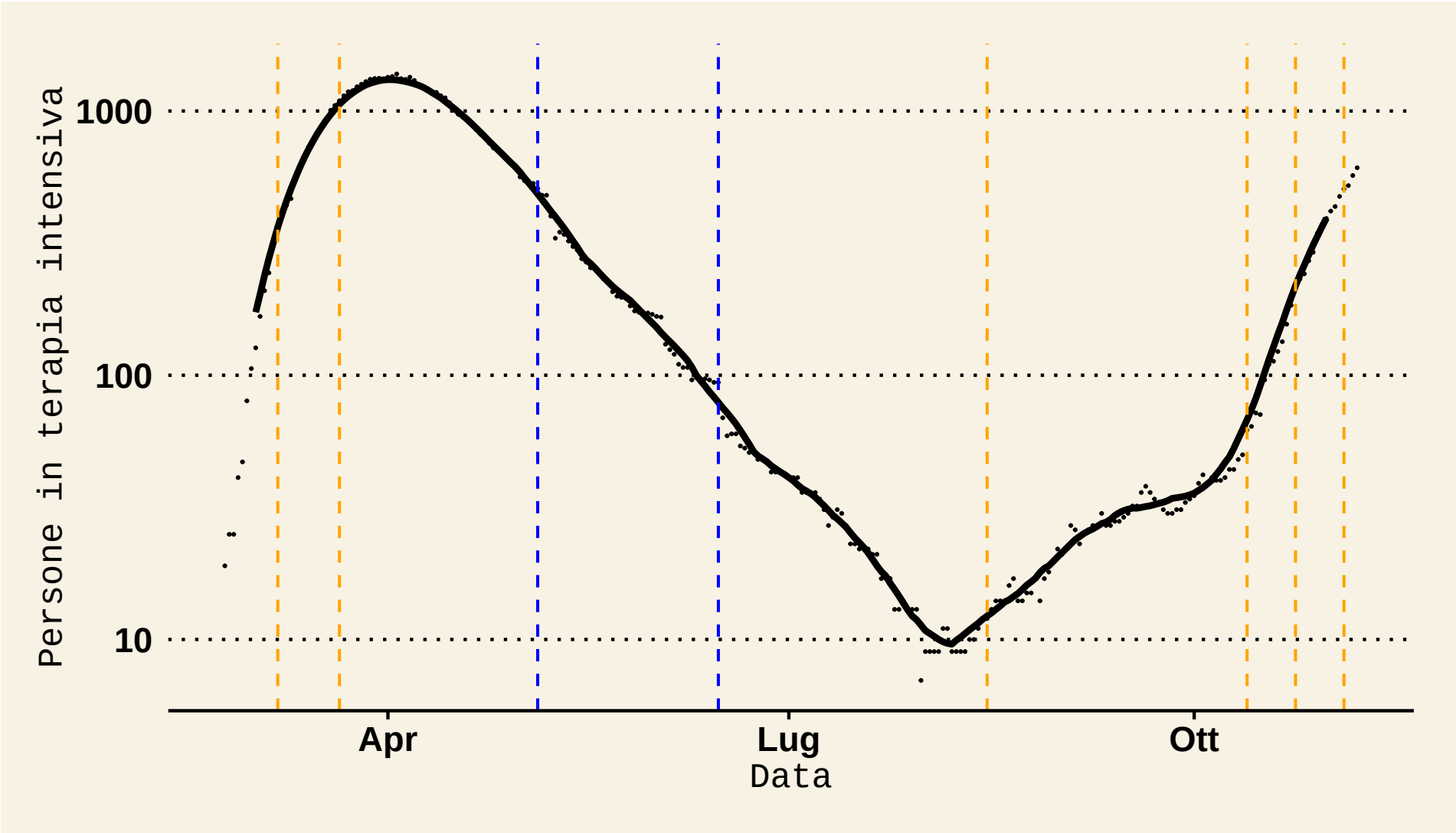
Numero di nuovi positivi (Lombardia)



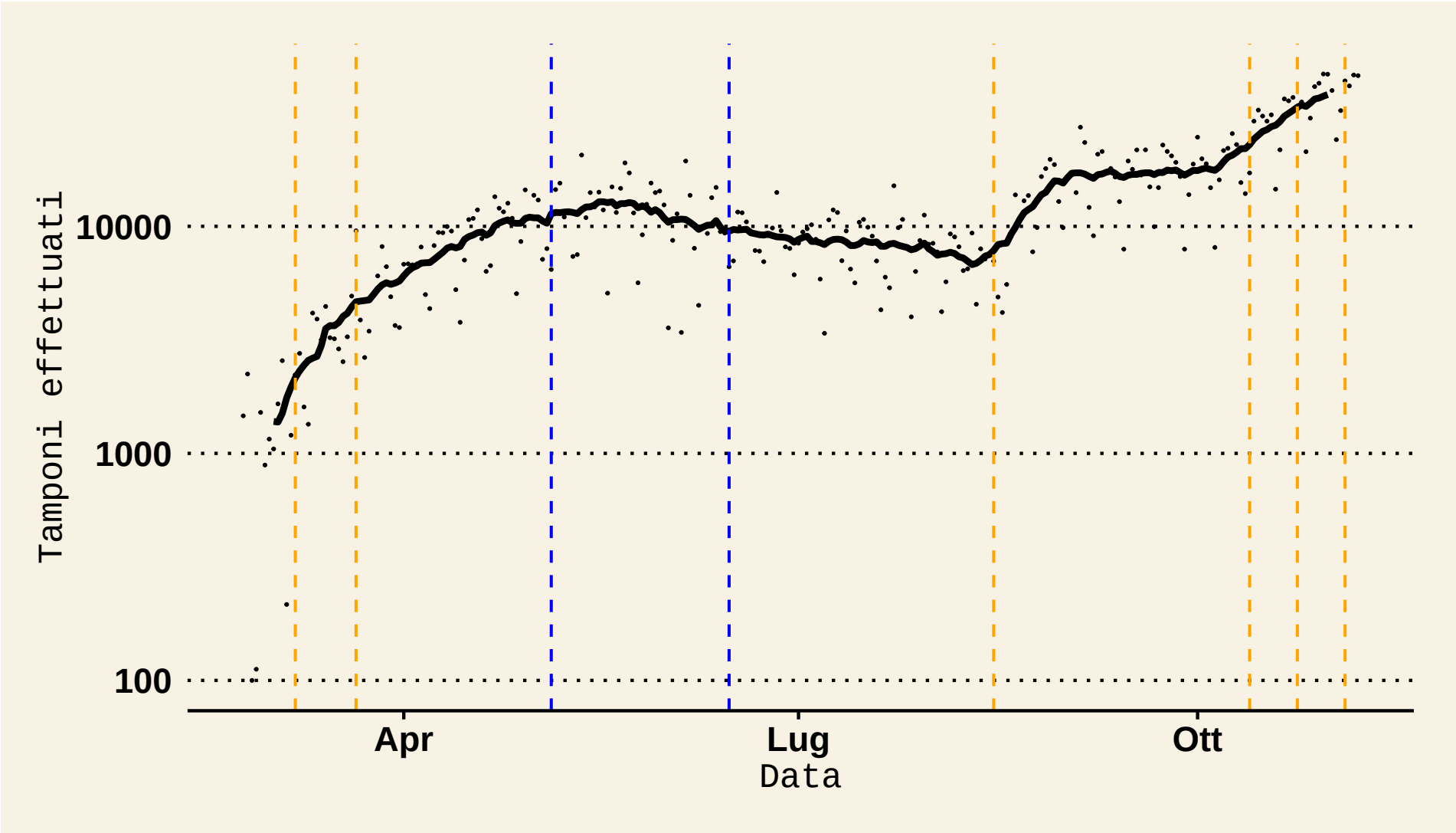
Numero di decessi (Lombardia)



Numero di pazienti in terapia intensiva (Lombardia)



Numero di tamponi effettuati (Lombardia)



Discussione dei risultati

- Tutti gli indicatori (numero di contagi, decessi, persone in terapia intensiva), suggeriscono che oggi 7 Novembre 2020 siamo in una **fase di forte crescita**.
- Il **primo picco** della pandemia è avvenuto, sulla base del numero dei nuovi positivi, intorno al 20-25 Marzo 2020.
- In estate, verso il 20-30 Luglio, si può osservare un **picco negativo**.
- È interessante notare come verso la fine di Settembre il contagio fosse in recessione in Lombardia, salvo poi risalire nuovamente e vertiginosamente ai primi di Ottobre.
- La data del secondo picco è purtroppo attualmente difficile se non impossibile da stimare.

Prevedere il contagio

- Il secondo problema che ci poniamo è di **prevedere** l'andamento del contagio.
- La previsione di lungo periodo è essenzialmente impossibile, dato che l'andamento dell'epidemia dipende dalle misure adottate dal governo stesso.
- Fare previsioni di **lungo periodo** è **poco sensato**. Possiamo però sperare di riuscire a fare previsioni nel **breve periodo** (= massimo 2 settimane).
- Anche in questo caso, è necessaria **enorme cautela**. Prevedere il passato (la stima del trend) è complesso, ma prevedere il futuro lo è molto di più.
- In questo caso, ci concentreremo sul numero di decessi e pazienti in terapia intensiva, i cui numeri sembrano essere più affidabili.

Un modello di regressione lineare semplice

- Sappiamo che **nella fase di crescita**, il corso dell'epidemia segue un andamento **esponenziale**.
- È quindi ragionevole assumere che

$$x_t = \gamma \lambda^t + \epsilon_t, \qquad t = 1, \ldots, T,$$

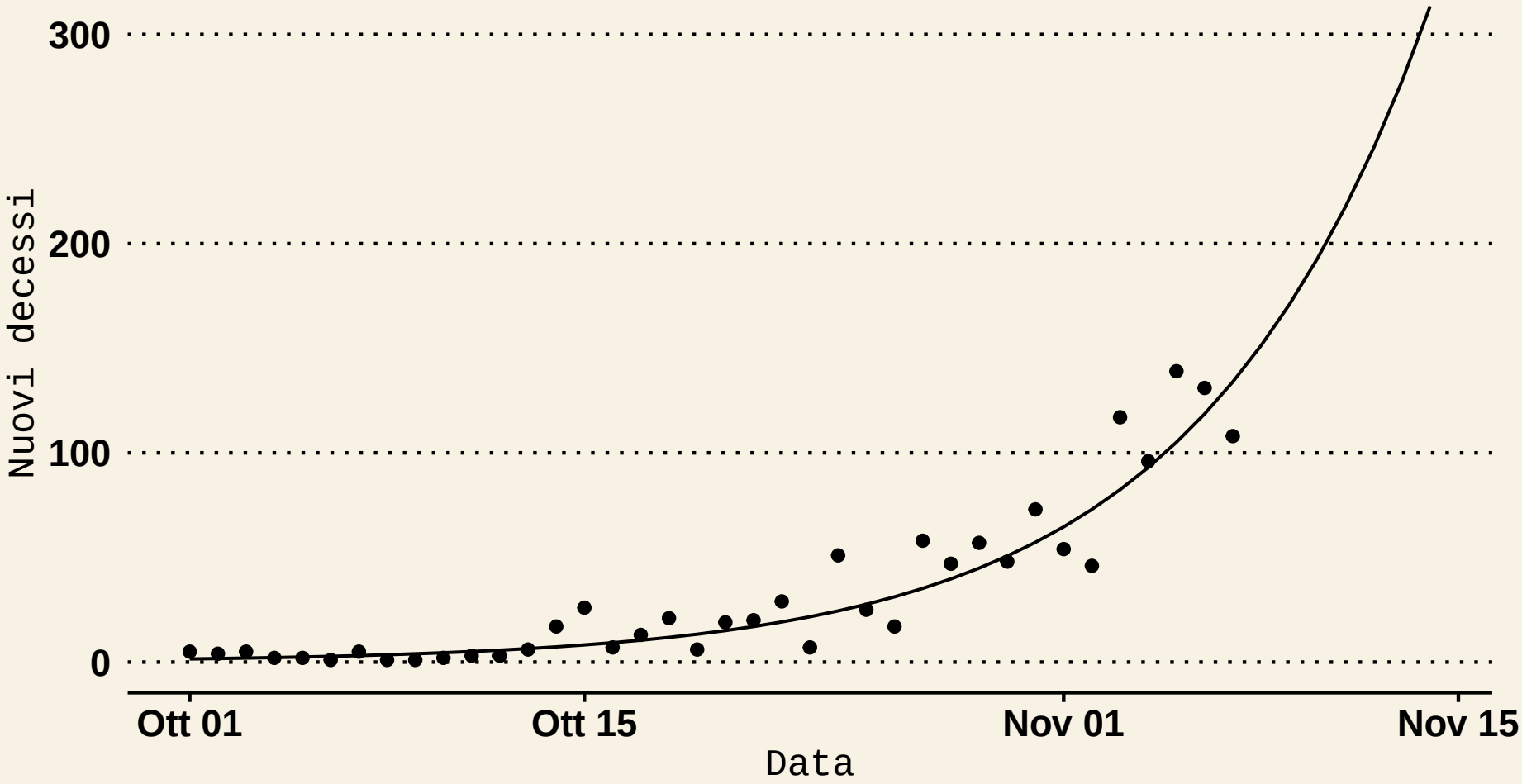
- Si tratta di un **modello linearizzabile**, infatti otteniamo che

$$\log x_t = \alpha + \beta t + \epsilon'_t, \qquad t = 1, \ldots, T.$$

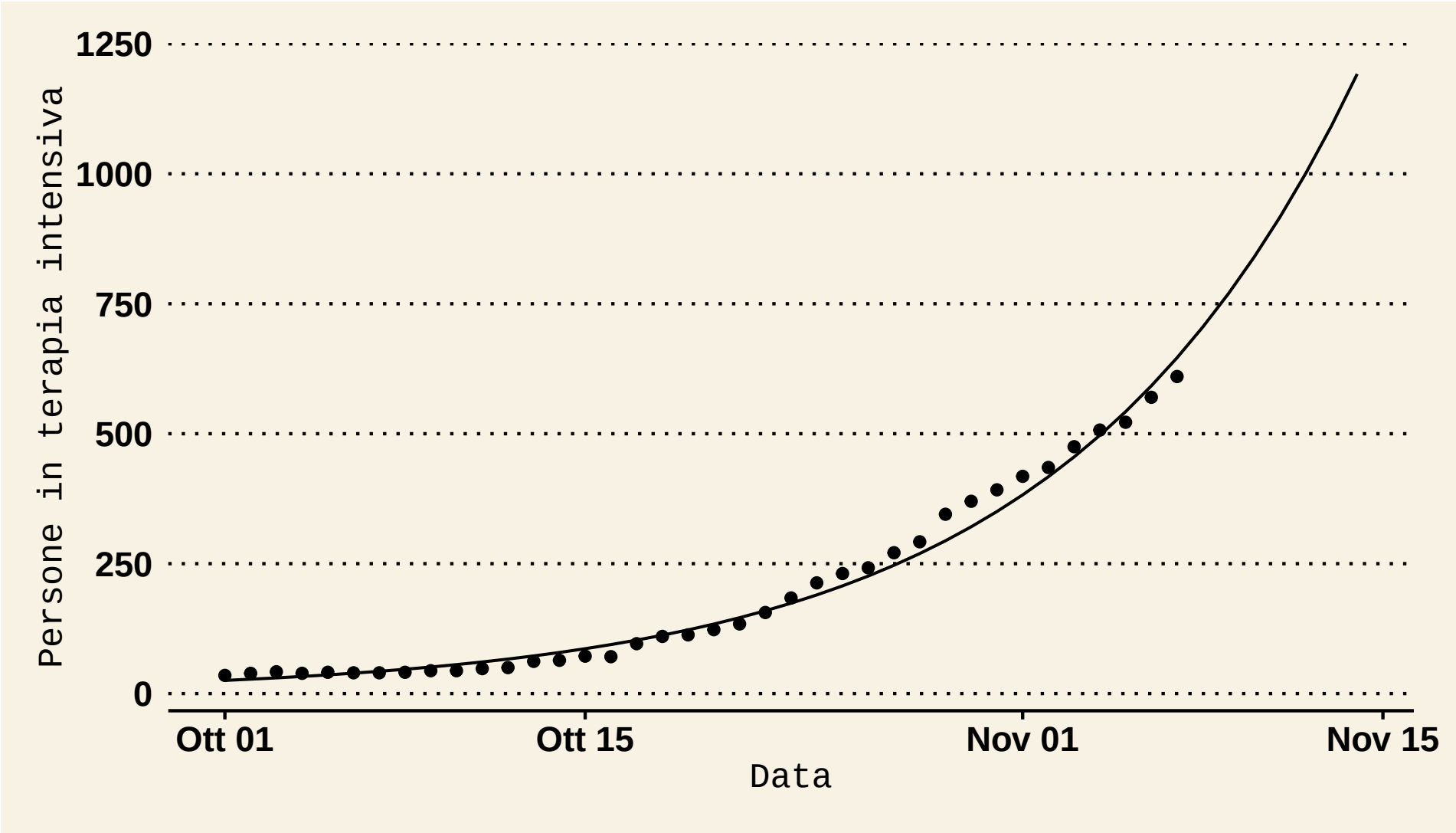
con $\alpha = \log \gamma$ e $\beta = \log \lambda$.

- Possiamo quindi stimare i parametri tramite il metodo dei minimi quadrati.

Previsione numero di decessi (Lombardia)



Previsione pazienti in terapia intensiva (Lombardia)



Discussione dei risultati

- Il modello esponenziale sembra essere (purtroppo) molto adeguato per la stima dei decessi. Viceversa, sembra essere un po' impreciso nel caso delle terapie intensive.
- Le linee continue rappresentano le previsioni per i prossimi 7 giorni.
- **Attenzione**. Si tratta di estrapolazioni, pertanto le previsioni sono affidabili solamente se il modello è valido.
- Il modello è sicuramente **falso** nel lungo periodo: la curva necessariamente raggiungerà un picco, mentre il modello esponenziale implica una crescita infinita.
- Il modello potrebbe essere inaffidabile anche nel breve periodo: ciò dipende dalle scelte del governo dei prossimi giorni e quelle che sono appena state prese.