Analizzando il plot di decomposizione della serie temporale con stagionalità multipla costruita considerando una possibile stagionalità settimanale, mensile e annuale dei dati, si osserva che le stagionalità settimanale e mensile sono in realtà piuttosto deboli. Si sceglie dunque di considerare unicamente la presenza di una stagionalità annuale (frequenza = 365).

Metodi utilizzati e background teorico:

1. **Decomposizione serie temporale, pacchetto *Anomalize* con metodo *Twitter*** →

<https://cran.r-project.org/web/packages/anomalize/vignettes/anomalize_quick_start_guide.html#further-understanding-methods>

<https://cran.r-project.org/web/packages/anomalize/vignettes/anomalize_methods.html>

Questo approccio si basa sulla seasonal decomposition di una serie temporale. Nello specifico, l’approccio si basa sull’analisi dei residui di una serie temporale che abbia rimosso sia le componenti stagionali che le componenti relative al trend. Il metodo usato per la decomposizione stagionale, che prende spunto da quello applicato nel pacchetto *AnomalyDetection* di Twitter, rimuove la componente legata al trend rimuovendo la mediana dei dati e funziona meglio del metodo di decomposizione STL nel caso di componente stagionale preponderante rispetto al trend, come nel caso dei dati in osservazione.

Per applicare il metodo di decomposizione, si definisce la frequenza come il numero di osservazioni che fanno parte di un ciclo stagionale nei dati (nel nostro caso 365 essendo la stagionalità annuale), mentre il trend è definito come il periodo temporale che può essere aggregato in modo da visualizzare la tendenza centrale dei dati (si decide, in questo senso, di lasciare l’impostazione di default per i dati annuali e dunque considerare un trend trimestrale). Si forniscono diverse possibilità (diverse rappresentazioni grafiche) per alpha e massima percentuale di outliers sul totale dei dati.

1. **Approccio ARIMA, pacchetto *tsoutliers*** →

<https://jalobe.com/doc/tsoutliers.pdf>

Considera diversi tipi di outliers: innovational outlier, additive outlier, level shift, temporary change and seasonal level shift. L’approccio consiste in tre fasi:

* Identificazione outliers: dato un modello ARIMA fittato sui dati, studio della significatività di tutti i tipi di outliers ai diversi time points, sfruttando la statistica test t nell’equazione di regressione dei residui stimati
* Rimozione outliers: la scelta iniziale del modello ARIMA potrebbe essere stata influenzata dalla presenza di outliers e vice versa. Si fitta quindi un modello ARIMA sulla serie originale che considera come regressori esterni gli effetti degli outliers potenziali identificati in precedenza. La significatività degli outliers è nuovamente analizzata nel nuovo modello e quelli non significativi sono rimossi dall’insieme degli outliers potenziali.
* Iterazione primi due step: dopo aver compiuto i due step precedenti per la serie originale, si iterano per la serie aggiustata; se si trovano nuovi outliers, sono aggiunti a quelli trovati nello step precedente. La serie aggiustata si ottiene sottraendo alla serie originale il total outlier effect, calcolato dividendo l’effetto dei diversi outliers per i coefficienti associati ai regressori esterni nel modello fittato. Dopo aver scelto e fittato un modello sulla serie aggiustata, se non ci sono più outliers il processo può considerarsi concluso.

1. **Metodo CART, isolation tree, pacchetto *isotree*** → <https://github.com/david-cortes/isotree>

<https://cran.r-project.org/web/packages/isotree/isotree.pdf>

L’approccio si basa sull'algoritmo Isolation Forest, il quale consiste nella suddivisione dei dati in sub-samples in base ad alcune colonne o attributi scelti casualmente. L’idea è che, più rara è un’osservazione, più alta è la probabilità che un random split uniforme ponga gli outliers da soli in un branch e meno saranno gli split necessari per raggiungere lo scopo. L’estensione di tale algoritmo consiste nella determinazione di una misura di distanza (o similarità) tra le osservazioni sulla base di quanti split siano necessari per separarle.

Nello specifico, il sistema implementato su R per il problema in analisi consiste nel fitting del modello sui dati giornalieri e, successivamente, sulla previsione relativa ai medesimi dati. Tale procedura consente di rilevare lo score di outlierness standardizzato, calcolato secondo una formula indicata sul paper originale di riferimento del pacchetto. Valori vicini a 1 indicano outliers forti, vicini a 0.5 outliers mediamente importanti e vicini a 0 valori prossimi alla normalità. Si sceglie di etichettare come outliers quei valori che presentano un valore di score superiore a 0.55, in modo da garantire un trade off tra numero di punti trovati e forza dell’anomalia riscontrata.