WaterFlow_Sector229

Thomas Houweling

10/8/2022

Intro

Per eseguire, cambia directory nella linea di codice nella cella in basso con la directory contenente il file 'flow (1).csv' e specifica se (i) multicore o no (Se si possiamo specificare modelli piu' complessi), (ii) salvare le immagini (5 o 6 per ciascun settore), (iii) salvare i dati (dati puliti, modello e forecast).

Librerie

Useremo soprattutto zoo, tsibble e fable. Purtroppo questi pacchetti hanno molte funzioni con nomi simili (viva la fantasia!), creando quindi problemi di masking. Pertanto, carichiamo solo due librerie e per il resto chiamiamo esplicitamente il pacchetto quando esequiamo una funzione ambigua. Prima, installiamo le librerie mancanti.

```
## Loading required package: ggplot2
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'tibble'
```

```
## Warning: replacing previous import 'lifecycle::last_warnings' by
## 'rlang::last_warnings' when loading 'pillar'
```

```
## Loading required package: dplyr
```

```
## Warning: package 'dplyr' was built under R version 4.0.5
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
## Loading required package: lubridate
```

```
## Warning: package 'lubridate' was built under R version 4.0.5
```

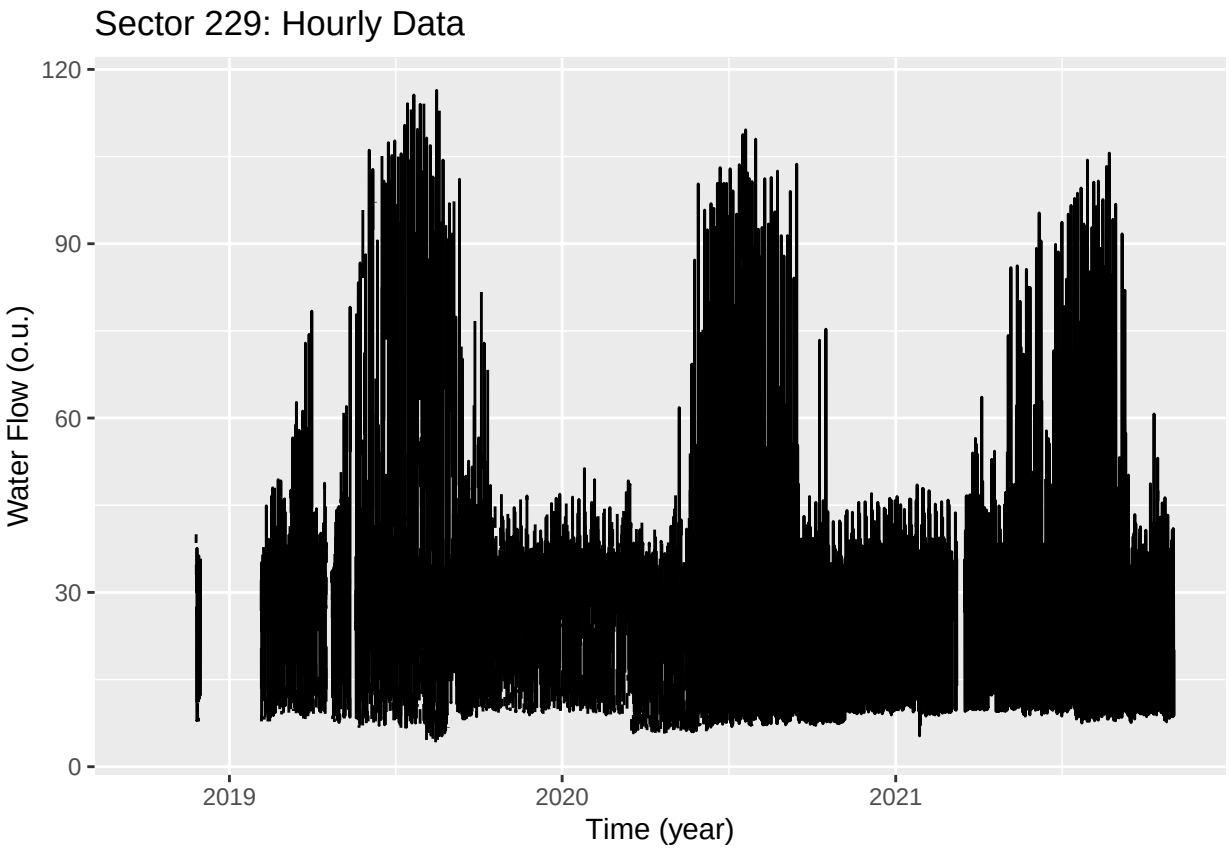
```
##
## Attaching package: 'lubridate'

## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union

## Loading required package: forecast

## Registered S3 method overwritten by 'quantmod':
##   method            from
##   as.zoo.data.frame zoo

## Loading required package: fable

## Warning: package 'fable' was built under R version 4.0.5

## Loading required package: fabletools

## Warning: package 'fabletools' was built under R version 4.0.5

##
## Attaching package: 'fabletools'

## The following objects are masked from 'package:forecast':
##
##     accuracy, forecast

## Loading required package: scales

## Loading required package: readr

## Warning: package 'readr' was built under R version 4.0.5

##
## Attaching package: 'readr'

## The following object is masked from 'package:scales':
##
##     col_factor
```

Carica dataset

In questo file esploriamo il primo settore nel dataset (settore 229) a titolo esemplificativo.

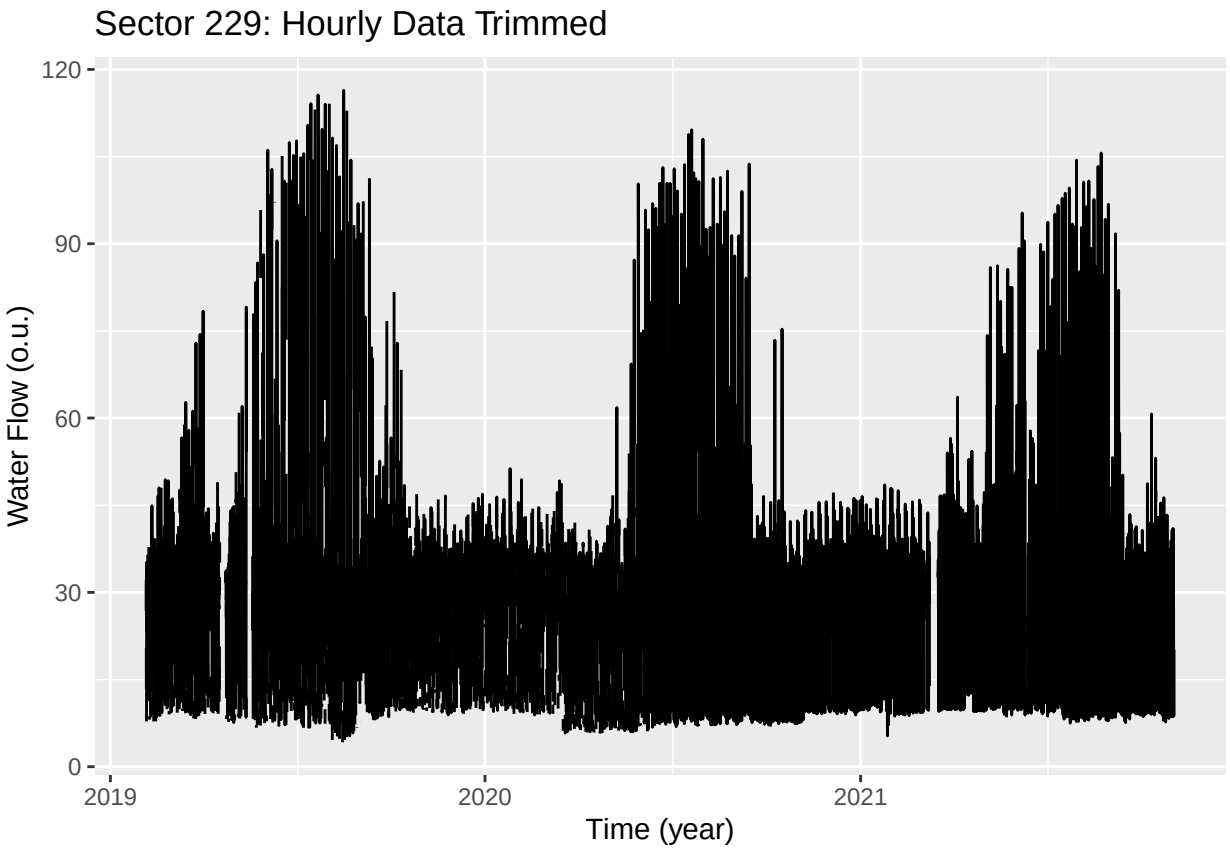
Regolarizzazione TS

I dati sono irregolari. Per creare time series (TS) regolari uso il pacchetto zoo.

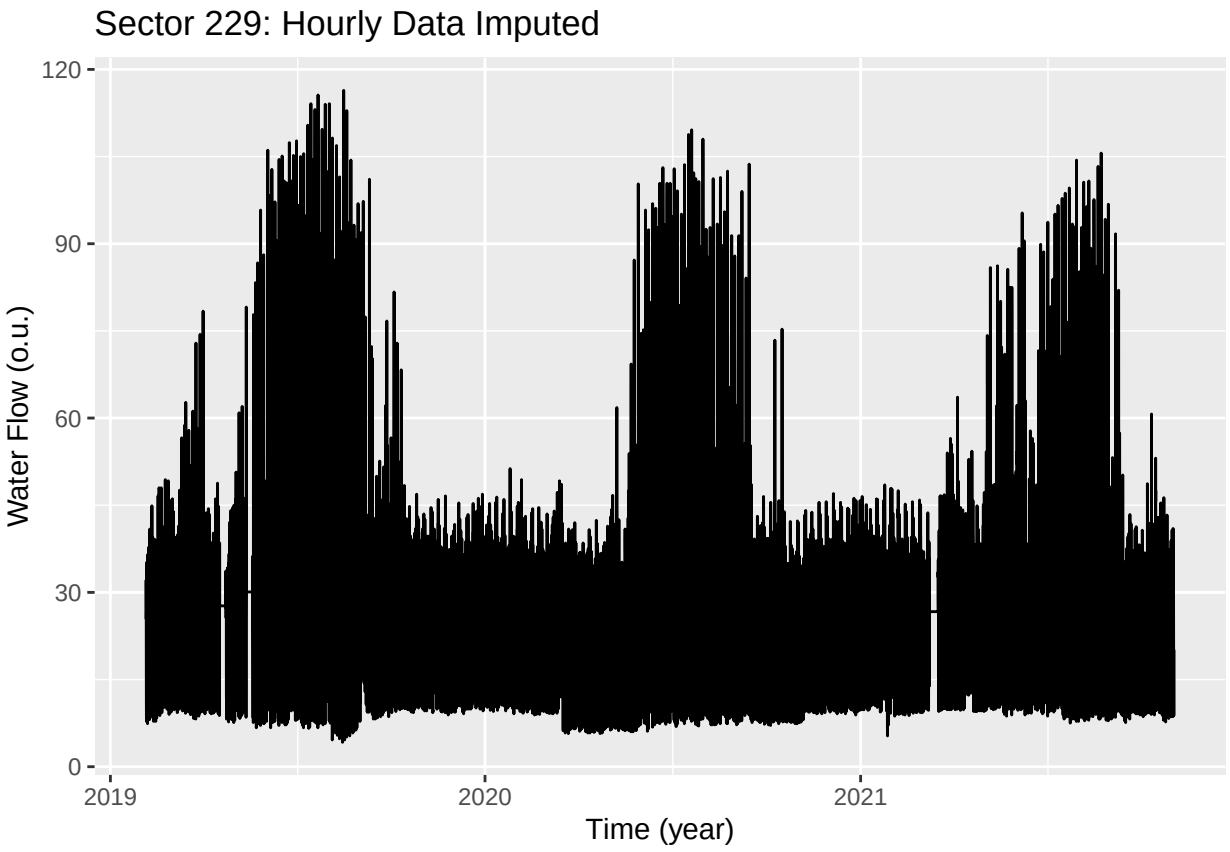


Pulizia dataset

Dall'immagine possiamo notare come vi siano larghi chunk di dati mancanti prima di Febbraio/Marzo 2019 e piccoli chunck di dati mancanti successivamente. Iniziamo rimuovendo i dati iniziali.



Molto meglio. Successivamente, interpoliamo i dati mancanti. Piuttosto che usare funzioni pre-compute (e.g., zoo::na.locf), scriviamo qualche linea di codice. Questo perche' i dati sono estremamente variabili (noisy) e pertanto e' meglio interpolare usando grandi chuncks di dati prima e dopo quelli mancanti. Specificamente, usiamo la media di tutti i valori nei 5 giorni antecedenti e successivi ai dati mancanti.



Data Inspection

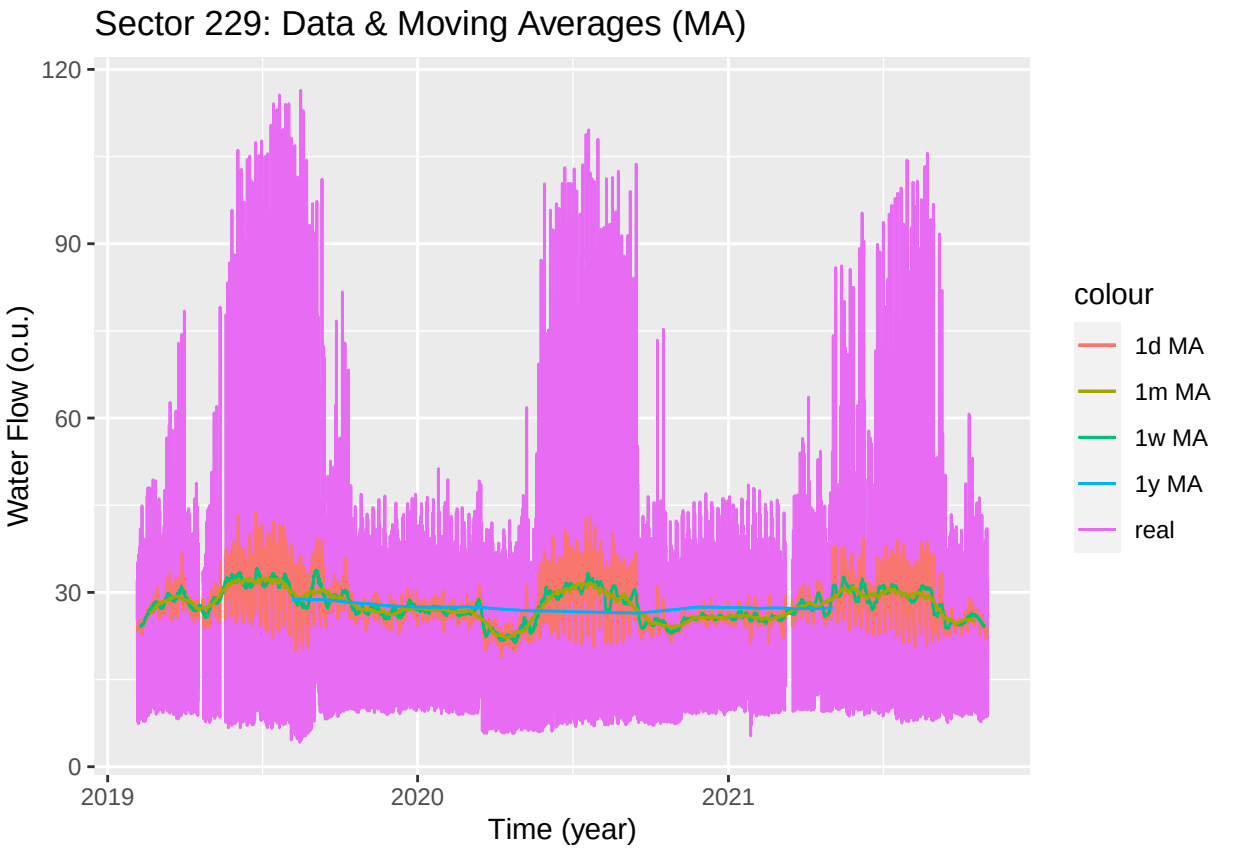
Dopo un minimo di pulizia iniziale, siamo pronti a ispezionare la nostra TS per stagionalita' e trends. La ciclicita' annuale e' apparente. Ci possono pero' essere anche stagionalita' settimanali, diurne, ecc... In piu' puo' esserci un long-term trend (lineare o non lineare). Visualizziamo delle moving averages per vedere se ci sono long-term trends.

```
## Warning: Removed 23 row(s) containing missing values (geom_path).
```

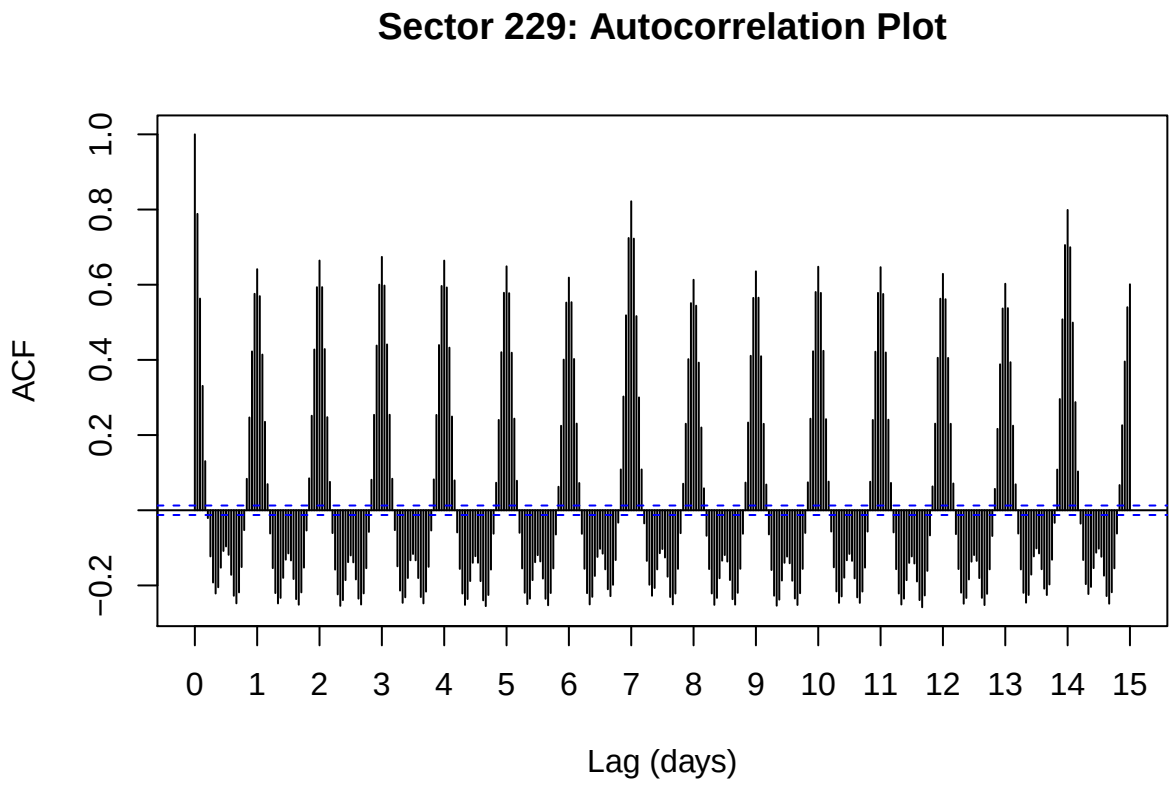
```
## Warning: Removed 167 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 719 row(s) containing missing values (geom_path).
```

```
## Warning: Removed 8759 row(s) containing missing values (geom_path).
```



Ispezionando le moving averages possiamo concludere che: ci e' una grande variabilita' intra-day, che supera di molto quella inter-day. Possiamo quindi aspettarci larghi intervalli di confidenza associati alla predizione. Le MAs di 1 settimana e 1 mese ci mostrano come la media di water flow in estate non sia realmente maggiore che di inverno di 10-20 unita' di misura, a fronte di un aumento spropositato nella variabilita' (varianza meno stazionaria della media). In piu', la moving average di periodo 1 anno e' piatta, suggerendo l'assenza di long-term trends. Ora procediamo a verificare la presenza di stagionalita' a periodo piu' breve (settimanale, diurno). Lo facciamo plottando il correlogramma a diversi lag.



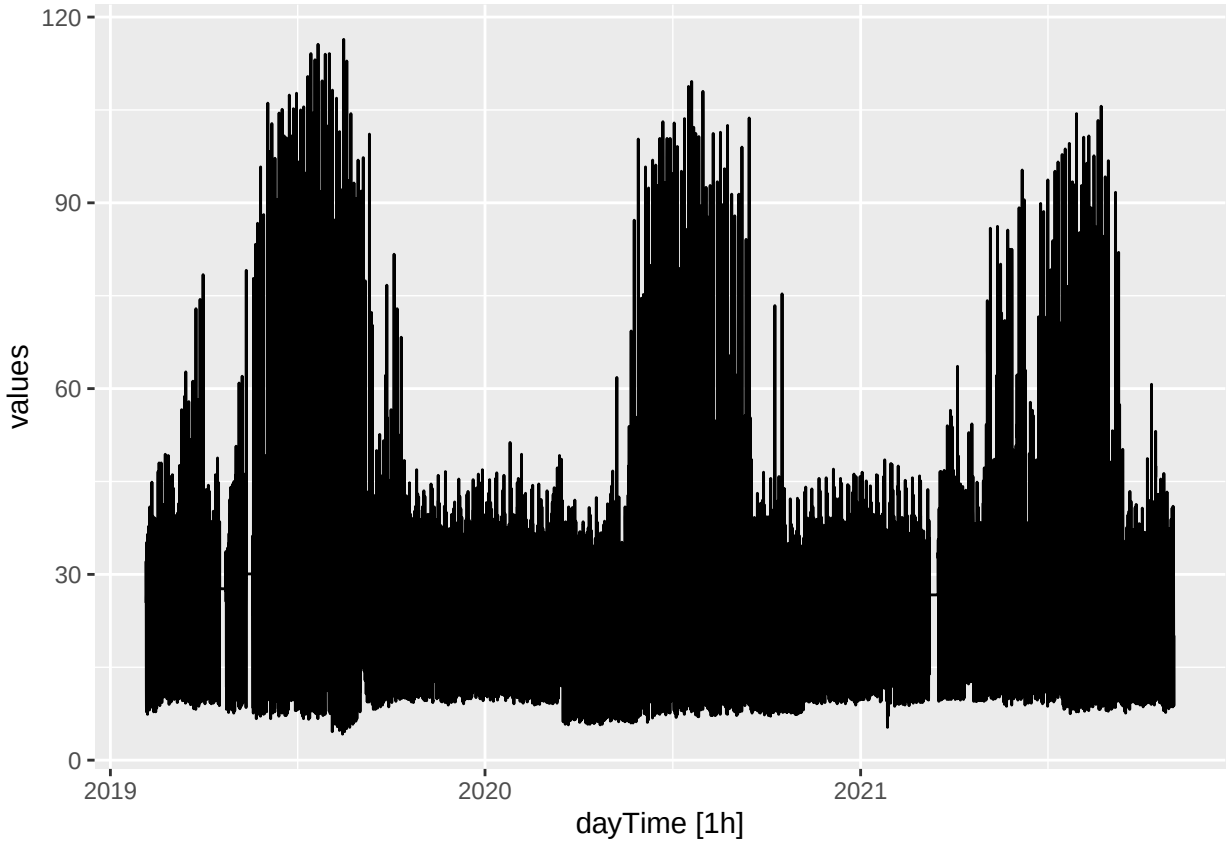
Un chiarissimo pattern giornaliero e' apprezzabile, insieme a un meno significativo pattern settimanale. Il modello che fitteremo ai dati deve quindi contenere almeno la stagionalita' annuale e giornaliera; meglio, se contiene anche quella settimanale. Procediamo ora a selezionare e fittare il modello, usando le librerie 'tsibble' e 'fable'.

Model fit

Con abbastanza risorse, e' possibile forzare un modello ARIMA con tutte e 3 le stagionalita' evidenziate precedentemente. Con le risorse a mia disposizione, riesco solo a fittare il modello di default che, fortunatamente, contiene le 2 maggiori stagionalita' (annuale e diurna). L'esecuzione di questa cella richiedera' un po' di tempo...

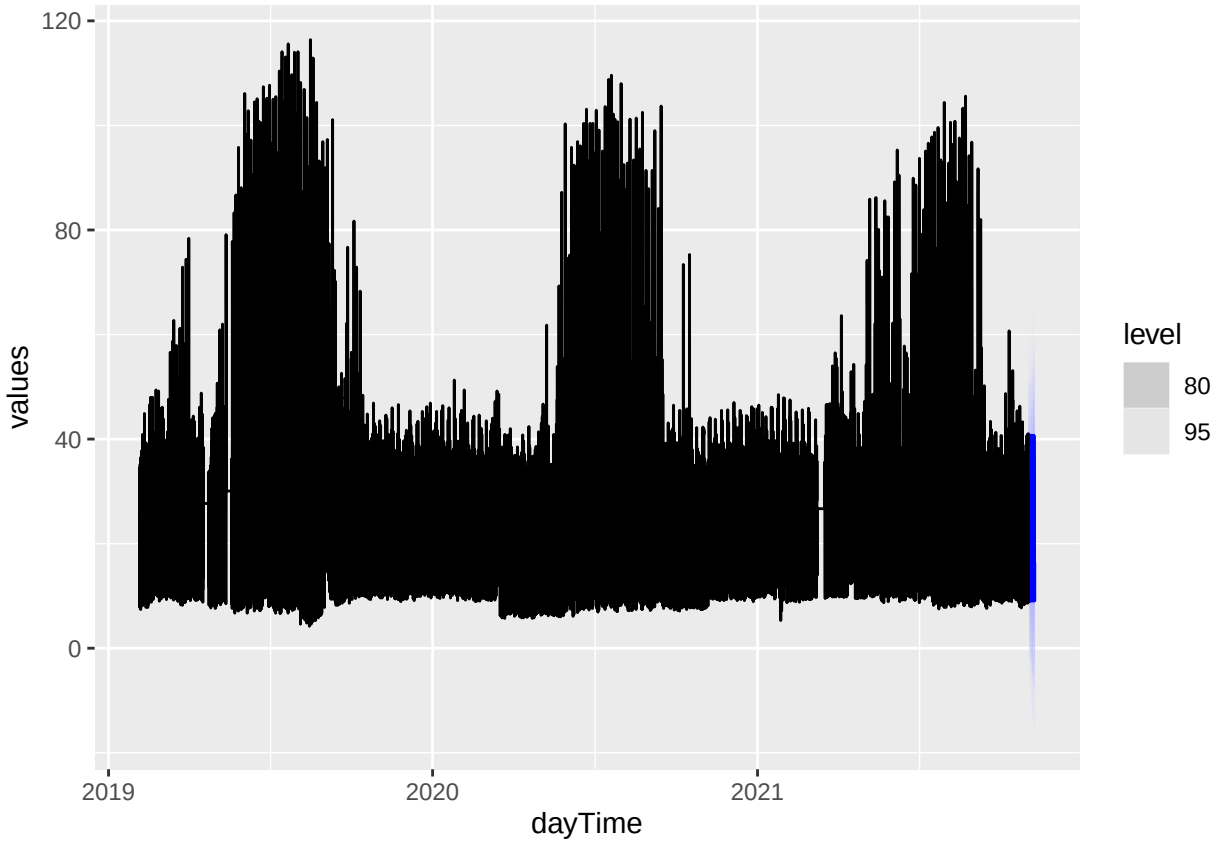
```
## Using `dayTime` as index variable.
```

```
## # A tibble: 1 x 1
##   .gaps
##   <lgl>
## 1 FALSE
```




```
## Series: values
## Model: ARIMA(1,0,2)(2,1,0)[24]
##
## Coefficients:
##          ar1      ma1     ma2     sar1     sar2
##       0.6106  -0.0711  0.0499  -0.7017  -0.3569
## s.e.  0.0135   0.0148  0.0102   0.0060   0.0060
##
## sigma^2 estimated as 48.55:  log likelihood=-80701.5
## AIC=161415   AICc=161415   BIC=161463.5
```

Forecast



Come si evince dagli intervalli di confidenza, le predizioni non sono estremamente accurate, coprendo un largo range di valori.