

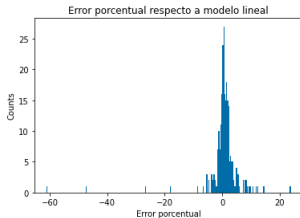
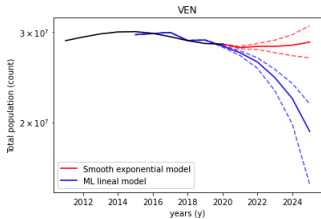
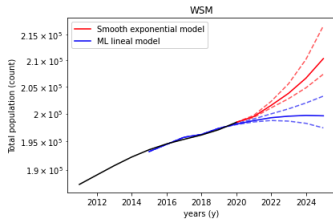
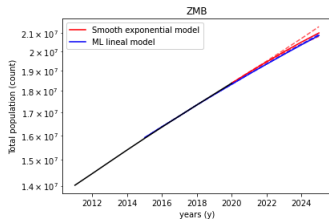
Predicción de la población para los próximos cuatro años

Test de programación – Jr. Data Engineer - **Ventagium**

F.E. Charry-Pastrana

2021.07.13

Resultados



Exploración de datos

Población P depende de tiempo t , natalidad n , mortalidad d y migración m .

$$\begin{aligned} P(t) &\propto f(t, n, d, m), \\ &\propto f(t, n(t), d(t), m(t)), \\ &\propto f(t, n(t), d(t)). \end{aligned}$$

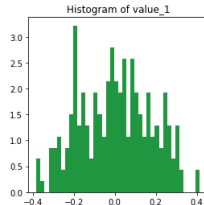
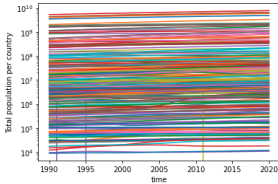
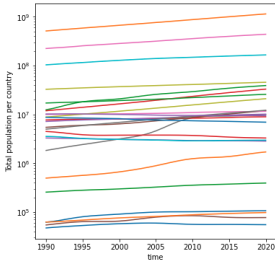
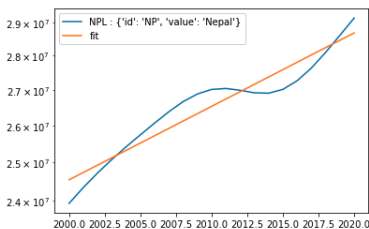
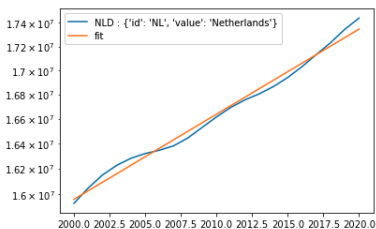


Figure: Correlación entre $P(t)$ y $m(t)$

$$\log(P(t)) = At + B$$

Modelos inicial: $P(t) = A \exp(B t)$

$P(t)$ depende únicamente del tiempo t de forma exponencial,



sin embargo, el modelo no captura las variaciones de los últimos años

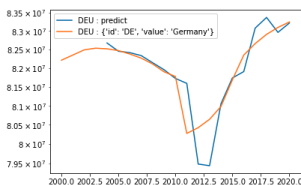
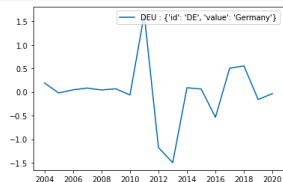
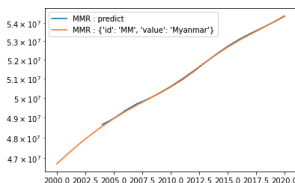
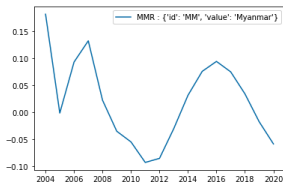
Modelos exponencial con corrección de los últimos años

$$P(t+1) = A \exp(Bt) + \bar{\delta},$$

$$\delta = P(t) - P_{\text{real}}(t)$$

```
y_predict = exp(x[j+4], parameters) - mean(delta[-2:])
```

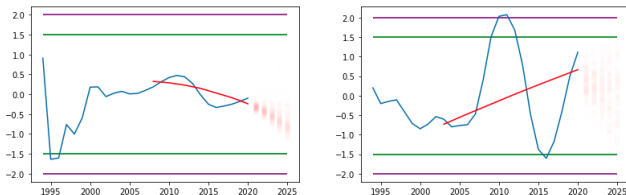
```
delta = y_predict - y[j+4]
```



¿Podemos predecir el «error»? Yes, yes, claro que yes.



Figure: Monte Carlo $\approx 10^4$ (10^2), value mínimo, máximo y promedio.



$P(t)$, $n(t)$ y $d(t)$ hasta $t \rightarrow 2024$.

ML: Lineal

- K-Nearest Neighbor regression (KNN)
- Support Vector Regression (SVR)
- Gaussian Processes (GP)

Modelo lineal utilizando t , $n(t)$ y $d(t)$.

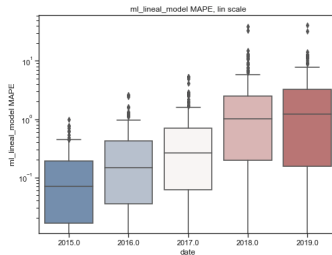
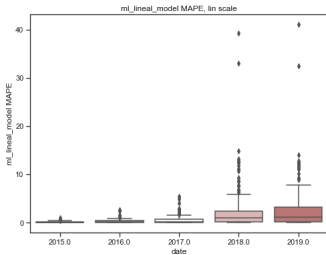
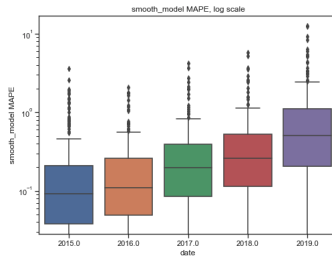
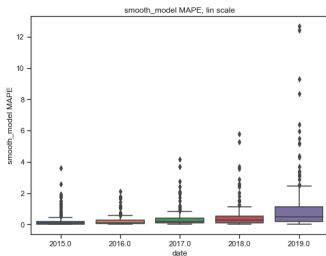
```
X = df_all[(df_all['date']<=2015) & (df_all['date']>=2010)]  
Y = df_all[(df_all['date']<=2015) & (df_all['date']>=2010)]
```

```
neigh = linear_model.Ridge(alpha=.5)
```

```
neigh.fit(X.values, Y.values)
```

```
neigh.predict(X_to_predict.values)
```

Errores: MAPE



Conclusiones

- Ambos modelos «lineales» nos permite predecir la población de cada país para los próximos cuatro años.
- Debido a la localidad del modelo «exponencial», se escoge el modelo lineal en referencia a los límites máximos, mínimos y promedio.
- Para el análisis de datos, es posible utilizar la pendiente del modelo inicial $P(t) = A \exp(B t)$ y sus correlaciones con otras variables (puede ser índices geográficos).
- Aunque falta realizar muchos experimentos y fine tune, estos modelos son un excelente primer paso en el entendimiento y la predicción de la población mundial. Se recomienda modelos avanzados una vez se entienda el fenómeno en modelos simples.