Tipologia i cicle de vida de les dades
Josep Queralt Fibla, Aïda Piñol Noguero

PRÀCTICA 2

1. Descripció del dataset. Perquè és important i quina pregunta/problema pretén respondre?

El daset que emprarem en aquesta pràctica prové del repositori de dades obertes de la FAO, el FAOSTAT, que depèn de la ONU i que recopila informació sobre la producció d'aliments arreu del món.

S'ha escollit els tres datasets següents:

1. Crop Production, Yield, Harvested Area (Global - National - Annual - FAOSTAT)

Dades sobre àrea (ha), producció (tonnes) i rendiment (hg/ha) a nivell nacional, global i anual.
Link: https://data.apps.fao.org/catalog/dataset/crop-production-yield-harvested-area-global-national-annual-faostat

2. Pesticides indicators (National - Global - Annual)

Dades sobre utilització de pesticides a nivell nacional, global i anual.
Link:
https://data.apps.fao.org/catalog/dataset/pesticidesindicatorsnationalglobalannualfaostat

3. Fertilizers indicators (National - Global - Annual)

Dades sobre utilització de fertilitzants a nivell nacional, global i anual.
Link:
https://data.apps.fao.org/catalog/dataset/fertilizers-indicators-national-global-annual-faostat

Amb aquestes dades, el que es pretén és donar resposta a les qüestions següents:

1. Com ha evolucionat la producció i l'ús de fertilitzants i pesticides?

2. Quin impacte té sobre la quantitat produïda?

2. Integració i selecció de les dades d'interès a analitzar. Pot ser el resultat d'addicionar diferents datasets o una subselecció útil de les dades originals, en base a l'objectiu que es vulgui aconseguir.

```
#-------------------------
# Selecció de les columnes da cada dataset
#-------------------------

df_c = df_c[['Area', 'Item', 'Element', 'Year', 'Value']]
df_c = df_c.pivot(index=['Area', 'Item', 'Year'], columns=['Element'])
df_c.reset_index(inplace=True)
df_c.columns = ['country', 'crop', 'year', 'area_harvested', 'production', 'yield']

# Com que les dades sobre producció tenen molt més detall del que necessitem fem un groupby i sumem
df_c = df_c.groupby(['year', 'country']).sum().reset_index()

df_f = df_f[['country_name_en', 'year', 'use_per_area_of_cropland_kg_ha']]
df_f = df_f.rename(columns={'country_name_en':'country','use_per_area_of_cropland_kg_ha':'nitrogen'})

df_p = df_p[['country_name_en', 'year', 'use_per_area_of_cropland_kg_ha']]
df_p = df_p.rename(columns={'country_name_en':'country', 'use_per_area_of_cropland_kg_ha':'pesticides'})
```

```
#-------------------------
# Integració dels diferents datasets
#-------------------------

df = df_c.merge(df_f, on=['country', 'year']).merge(df_p, on=['country', 'year'])

#-------------------------
# Selecció només de països
#-------------------------

# Eliminem no països

nopais = ['World', 'Americas', 'South America', 'Asia',
 'Northern America',
       'Net Food Importing Developing Countries', 'Europe',
       'European Union (28)', 'European Union (27)', 'Southern Europe',
       'Southern Asia', 'Africa', 'Low Income Food Deficit Countries',
       'Central America', 'Northern Africa',
       'Eastern Asia', 'Western Asia',
       'China, mainland', 'South-eastern Asia', 'Southern Africa',
        'Least Developed Countries',
        'Small Island Developing States',
       'Land Locked Developing Countries', 'Caribbean',
        'Western Africa', 'Oceania','Eastern Africa']


dfp = df[~df['country'].isin(nopais)]
```

3. Neteja de les dades.

3.1. Les dades contenen zeros o elements buits? Gestiona cadascun d'aquests casos.

```
#-------------------------
# Neteja de les dades
#-------------------------

# Valors nuls
df.isnull().sum()

#Zeros
df.eq(0).sum()
```

3.2. Identifica i gestiona els valors extrems.

```
# Outliers

dfp_outliers = dfp[['country', 'area_harvested', 'production', 'yield', 'nitrogen', 'pesticides']]

# Calculem el IQR per cada columna
Q1 = dfp_outliers.quantile(0.25)
Q3 = dfp_outliers.quantile(0.75)
IQR = Q3 - Q1

# Trobem els valors que estan fora del rang interquartílic
outliers = (dfp_outliers < (Q1 - 1.5 * IQR)) | (dfp_outliers > (Q3 + 1.5 * IQR))

# Seleccionem els outliers
outlier_rows = dfp_outliers[outliers.any(axis=1)]
```

4. Anàlisi de les dades.

4.1. Selecció dels grups de dades que es volen analitzar/comparar (p. e., si es volen comparar grups de dades, quins són aquests grups i quins tipus d'anàlisi s'aplicaran?).

El nostre dataset són dades de panell. Tenim una observació per any i país amb la producció total agrícola, el terreny utilitzat, la productivitat i la quantitat de nitròg en i de pesticides emprada. D'acord amb les dades, les comparacions que es volen fer seran entre els diferents països.

4.2. Comprovació de la normalitat i homogeneïtat de la variància.

La comprovació de la normalitat es farà sobre cadascuna de les variables del dataset i la homogeneïtat de la variància es farà comparant-les entre elles.

Primer comprovarem la normalitat de cadascuna de les variables del nostre dataset amb el següent codi:

```
#-------------------------
# Test de normalitat
#-------------------------

for column in df[['area_harvested', 'production', 'yield', 'nitrogen', 'pesticides']].columns:
    stat, p_value = normaltest(df[column])
    alpha = 0.05

    print(f"Column: {column}")
    print(f"Test statistic: {stat}")
    print(f"P-value: {p_value}")

    if p_value > alpha:
        print("Les dades es distribueixen normalment.\n")
    else:
        print("Les dades no es distribueixen normalment.\n")
```

I obtenim el següent resultat:

```
Column: area_harvested
Test statistic: 3146.283548386153
P-value: 0.0
Les dades no es distribueixen normalment.

Column: production
Test statistic: 3445.7563929115017
P-value: 0.0
Les dades no es distribueixen normalment.

Column: yield
Test statistic: 3187.9990013208144
P-value: 0.0
Les dades no es distribueixen normalment.

Column: nitrogen
Test statistic: 1396.551322559501
P-value: 5.530106038241989e-304
Les dades no es distribueixen normalment.

Column: pesticides
Test statistic: 1592.0894384270373
P-value: 0.0
Les dades no es distribueixen normalment.
```

Seguidament, amb aquest codi realitzem el test de Levene per comprovar l'homoscedasticitat de les nostres dades:

```
#-------------------------
# Test d'homocedasticitat
#-------------------------

stat, p_value = levene(df['area_harvested'], df['production'], df['yield'],
                       df['nitrogen'], df['pesticides'], df['yield'])
alpha = 0.05

print("Test d'homoscedasticitat")
print(f"Test statistic: {stat}")
print(f"P-value: {p_value}")

if p_value > alpha:
    print("Les dades mostren homoscedasticitat.\n")
else:
    print("Les dades no mostren homoscedasticitat.\n")
```

I, tal com veiem a l'output, veiem que les dades no són homocedàstiques:

```
Test d'homoscedasticitat
Test statistic: 203.68421815613158
P-value: 9.449606755315247e-212
Les dades no mostren homoscedasticitat.
```

4.3. Aplicació de proves estadístiques per comparar els grups de dades. En funció de les dades i de l'objectiu de l'estudi, aplicar proves de contrast d'hipòtesis, correlacions, regressions, etc. Aplicar almenys tres mètodes d'anàlisi diferents.

Primer s'aplicarà un test de correlació entre la *productivitat* i la *quantitat de nitrogen utilitzada* i la *productivitat i la quantitat de pesticida emprada*. Per fer-ho, es generarà un subset del nostre dataset per quedar-nos només amb les dades d'un sol país, en aquest cas, la Xina.

```
#------------------------
# Test de correlació
#------------------------

df_xina = df[df.country == 'China']

corr, p_value = pearsonr(df_xina['yield'], df_xina['nitrogen'])

print(f"Correlation coefficient: {corr}")
print(f"P-value: {p_value}")

if p_value < 0.05:
    print("Hi ha una correlació significativa.\n")
else:
    print("No hi ha una correlació significativa.\n")

corr, p_value = pearsonr(df_xina['yield'], df_xina['pesticides'])

print(f"Correlation coefficient: {corr}")
print(f"P-value: {p_value}")

if p_value < 0.05:
    print("Hi ha una correlació significativa.\n")
else:
    print("No hi ha una correlació significativa.\n")
```

En ambdós casos obtenim una correlació significativa entre la productivitat i el nitrogen per una banda, i entre la productivitat i l'ús de pesticides, per l'altra.

```
Correlation coefficient: 0.7630588778414624
P-value: 9.107442822772446e-05
Hi ha una correlació significativa.

Correlation coefficient: 0.9331231423460058
P-value: 1.9905914480113555e-09
Hi ha una correlació significativa.
```

La ***prova ADF*** [1] s'utilitza per avaluar la estacionarietat d'una sèrie temporal, que és una suposició clau en molts models de sèries temporals. Prova si hi ha una arrel unitària present en el model autorregressiu de les dades. Una sèrie temporal estacionària té mitjana, variància i estructura d'autocovariància constants al llarg del temps.

[1] QuantSpace. *Analizando Series Temporales con Python*. https://quantspace.es/2020/08/01/analisis-de-series-temporales-con-python-parte-2/ [Enllaç consultat el 14/06/2023].

```
#-------------------------
# Test ADF
#-------------------------

for column in df_xina[['area_harvested', 'production', 'yield', 'nitrogen', 'pesticides']].columns:
    result = adfuller(df_xina[column])

    test_statistic = result[0]
    p_value = result[1]

    print(f"Column: {column}")
    print(f"Test Statistic: {test_statistic}")
    print(f"P-value: {p_value}")

    if p_value < 0.05:
        print("És estacionària.\n")
    else:
        print("No és estacionària.\n")
```

El resultat obtingut evidencia que les series per a Xina no són estacionàries:

```
Column: area_harvested
Test Statistic: -0.23523605609860776
P-value: 0.9341891608812708
No és estacionària.

Column: production
Test Statistic: -1.2092755485948279
P-value: 0.6695929314488717
No és estacionària.

Column: yield
Test Statistic: 4.653974670860663
P-value: 1.0
No és estacionària.

Column: nitrogen
Test Statistic: -1.729375516913094
P-value: 0.4160443978730236
No és estacionària.

Column: pesticides
Test Statistic: -2.020065840051365
P-value: 0.2778448178038425
No és estacionària.
```

Finalment, practicarem un ***regressió fixed-effects*** [2]sobre el conjunt de dades. La variable dependent serà la productivitat de la terra (yield) i les variables independents el *nitrogen* i els *pesticides*.

```
#-------------------------
# Regressió
#-------------------------

model = sm.OLS(df['yield'], sm.add_constant(df[['nitrogen', 'pesticides']]))
fixed_effects_model = model.fit(cov_type='cluster', cov_kwds={'groups': df['country']})

print(fixed_effects_model.summary())
```

Com veiem en l'output de la regressió, l'efecte d'ambdós regressors és positiu però només significatiu amb un nivell superior al 5% en el cas del nitrogen. Amb aquest resultat podem afirmar que l'administració de nitrogen té un efecte beneficiós en la productivitat de la terra.

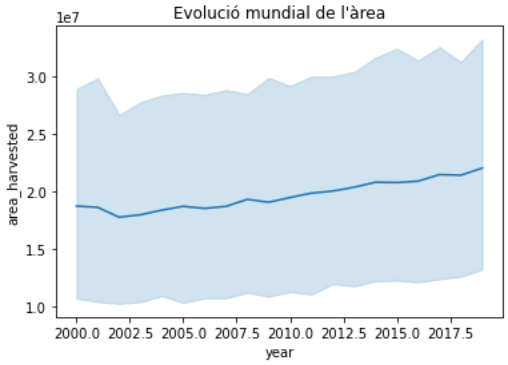
[2] Econometrics with r. *Fixed Effects Regression*. https://www.econometrics-with-r.org/10-3-fixed-effects-regression.html [Enllaç consultat el 14/06/2023].

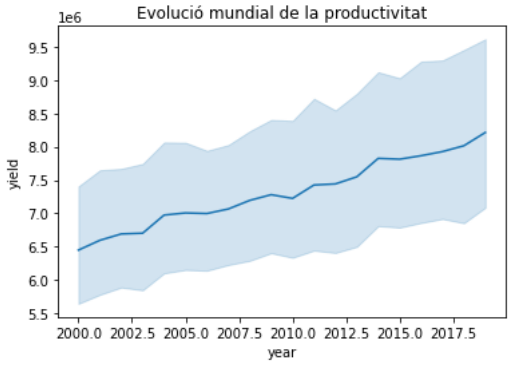
```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  yield   R-squared:                       0.210
Model:                            OLS   Adj. R-squared:                  0.209
Method:                 Least Squares   F-statistic:                     4.792
Date:                Fri, 16 Jun 2023   Prob (F-statistic):            0.00962
Time:                        18:46:23   Log-Likelihood:                -49424.
No. Observations:                2905   AIC:                         9.885e+04
Df Residuals:                    2902   BIC:                         9.887e+04
Df Model:                           2
Covariance Type:              cluster
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const       4.353e+06   6.75e+05      6.445      0.000    3.03e+06    5.68e+06
nitrogen    4.284e+04   1.65e+04      2.603      0.009    1.06e+04    7.51e+04
pesticides  1.397e+05   9.56e+04      1.462      0.144   -4.76e+04    3.27e+05
==============================================================================
Omnibus:                     3044.817   Durbin-Watson:                   2.105
Prob(Omnibus):                  0.000   Jarque-Bera (JB):           276624.796
Skew:                           5.067   Prob(JB):                         0.00
Kurtosis:                      49.719   Cond. No.                         122.
==============================================================================

Notes:
[1] Standard Errors are robust to cluster correlation (cluster)
```

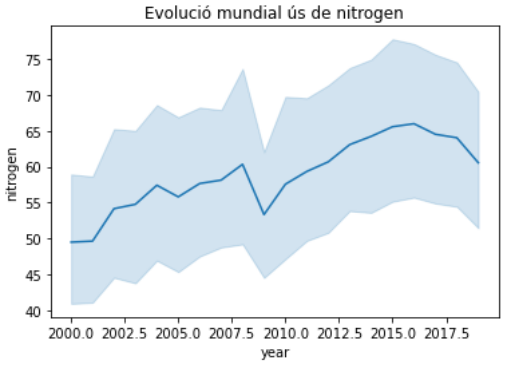
5. Representació dels resultats:

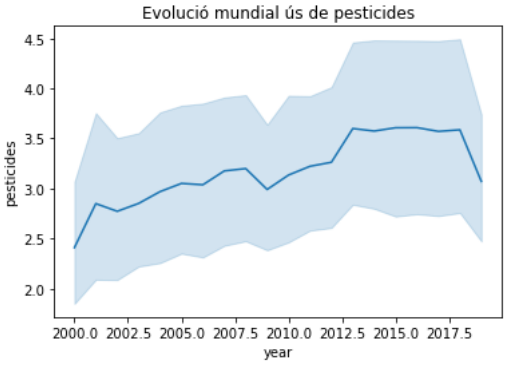
En els següents gràfics hem fet una anàlisi exploratòria de les dades. En els primers quatre es pot comprovar com tant l'àrea com la productivitat mostren una tendència positiva.



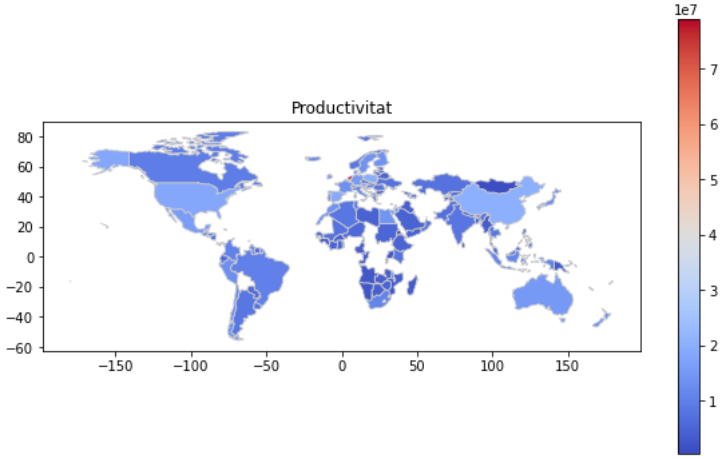


En canvi, el consum de nitrogen i de pesticides sembla haver experimentat un retrocés en els darrers anys.

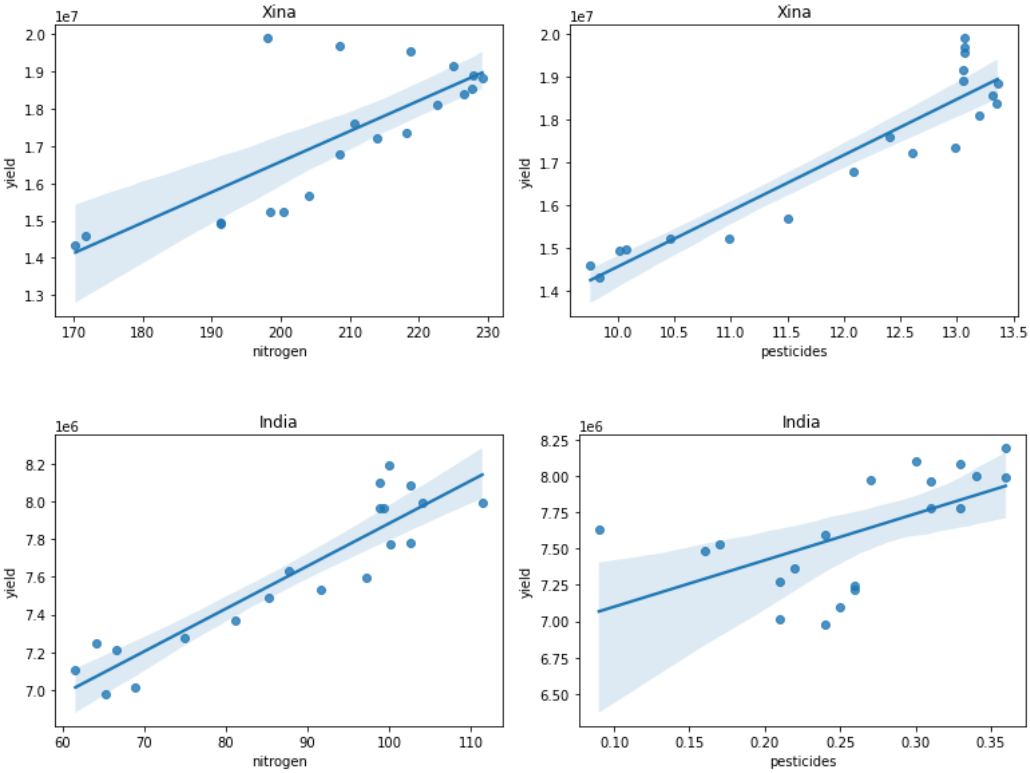


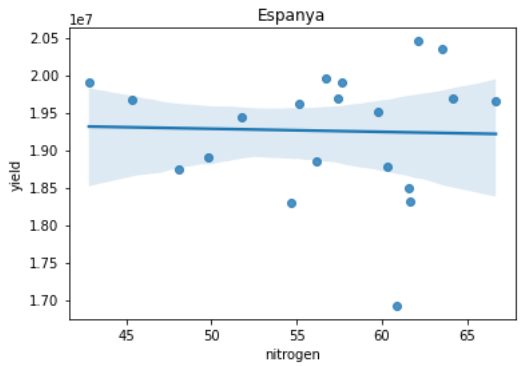


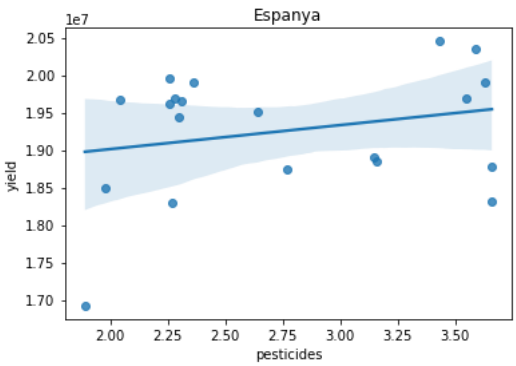
El mapa ens mostra, per als països on tenim dades, quina és la seva productivitat.



Finalment, hem fet uns gràfics sobre la relació entre el consum de nitrogen i pesticides i la productivitat a diferents països. En els casos xinès i indi veiem com un major ús d'aquests ensums es tradueix en una major productivitat de la terra, mentre que pel cas espanyol no es pot extreure la mateixa conclusió.







6. Resolució del problema:

A tal de conclusió, davant les preguntes objecte d'anàlisis: *1. Com ha evolucionat la producció i l'ús de fertilitzants i pesticides?* I *2. Quin impacte té sobre la quantitat produïda?, trobem que* l'administració de nitrogen té un efecte beneficiós en la productivitat de la terra. Més concretament, els resultats han ressaltat que a nivell mundial, hi ha una tendència a l'alça de l'àrea i la productivitat i una disminució en el consum de nitrogen i de pesticides.

Respecte als casos concrets analitzats, tant l'ús de nitrogen com de pesticides a la Xina i a la Índia es correlacionarien en una major productivitat de la terra, mentre que a Espanya, no s'observaria aquesta relació.

7. Enllaç práctica: https://github.com/josquefi/pra_2