Progetto I

Simone Manti, Matricola: 566908

21/11/2021

Introduzione

La seguente analisi è rivolta ad industrie agricole, precisamente a quelle interessate ai fertilizzanti ed alle colture foraggere. L'obiettivo è capire quanto l'uso di fertilizzanti in alcune colture influenza le regioni italiane, in modo che queste possano scegliere il meglio per loro con meno componenti principali possibili.

Presentazione del problema

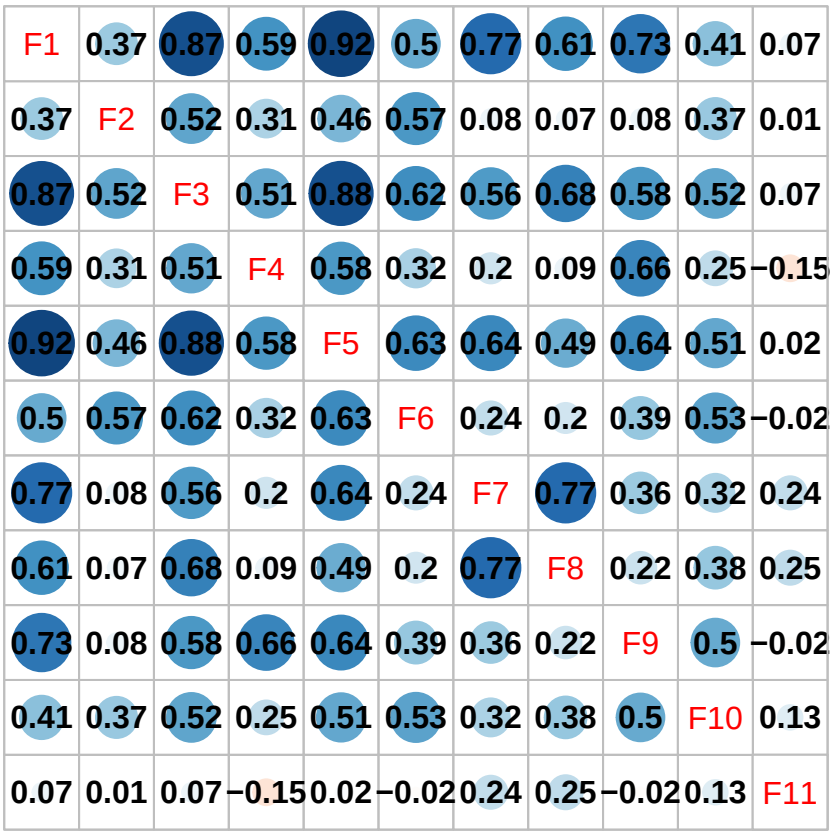
I dati esaminati risalgono al 2019 e provengono da tre diverse tabelle fornite dall'Istat (dettagli in fondo alla relazione), abbiamo scelto di non considerare alcuni dei fattori presenti perchè fortemente correlati ad altri. Inoltre anzichè considerare Toscana, Umbria, Marche e Lazio come osservazioni distinte, le abbiamo raggruppate nell'unica osservazione "Centro" (per alcuni fattori avevamo solo dati "in blocco").

I parametri presi in considerazione sono i seguenti:

- F1 Concimi minerali semplici
- F2 Concimi minerali a base di meso-elementi e di micro-elementi
- F3 Concimi minerali compositi
- F4 Concimi organici
- F5 Concimi organo-minerali
- F6 Prodotti fitosanitari e principi attivi vari
- F7 Mais ceroso
- F8 Loietto
- F9 Erba medica
- F10 Prati permanenti
- F11 Altri erbai monofiti

I fattori da F1 a F5 sono stati misurati in tonnellate e i dati si riferiscono a materiali solidi, mentre per i restanti fattori i dati si riferiscono alla superficie totale misurata in ettari.

Iniziamo la nostra analisi studiando la matrice delle correlazioni



Si possono già fare alcune considerazioni

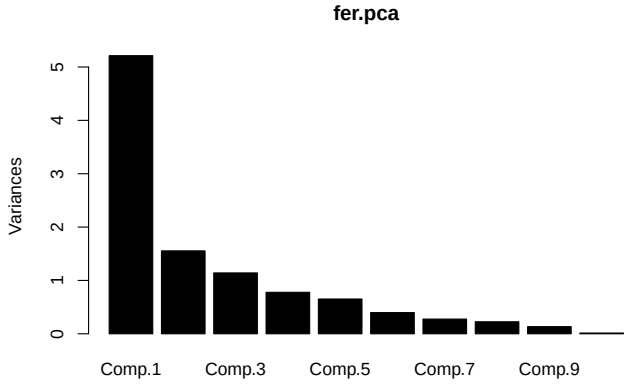
- F1 (concimi minerali semplici) è fortemente correlato con F3(concimi minerali compositi), F5(concimi organo minerali), F7(superficie in ettari di mais ceroso) e F9(superficie in ettari di erba medica)
- Il resto dei fattori non è molto correlato: questo rende l'analisi molto interessante seppur a priori difficile da interpretare.

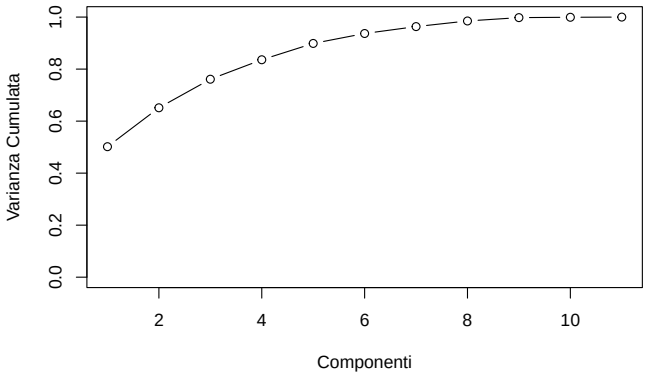
Analisi delle componenti principali

Dopo aver opportunamente standardizzato i dati, studiamo preliminarmente le componenti principali e vediamo un primo risultato riassuntivo.

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3     Comp.4     Comp.5
## Standard deviation     2.2830023 1.2465434 1.0680846 0.88183894 0.80731952
## Proportion of Variance 0.5016994 0.1495704 0.1098101 0.07485304 0.06273672
## Cumulative Proportion  0.5016994 0.6512698 0.7610799 0.83593295 0.89866967
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.62999435 0.52493461 0.47454782 0.36564374 0.11928334
## Proportion of Variance 0.03820359 0.02652414 0.02167658 0.01286907 0.00136959
## Cumulative Proportion  0.93687326 0.96339740 0.98507399 0.99794306 0.99931265
##                           Comp.11
## Standard deviation     0.0845033412
## Proportion of Variance 0.0006873511
## Cumulative Proportion  1.0000000000
```

Notiamo come le prime 3 componenti riescono a descrivere il 76% di varianza spiegata e con 4 arriviamo all'83%. Per maggiore chiarezza, riportiamo anche i plot delle varianze spiegate e della varianza cumulata rispetto alle componenti:



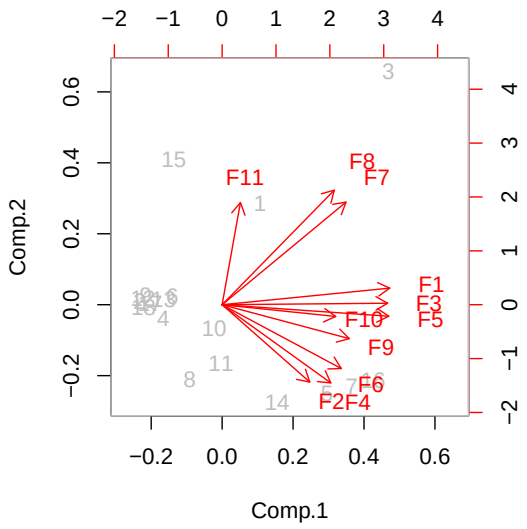


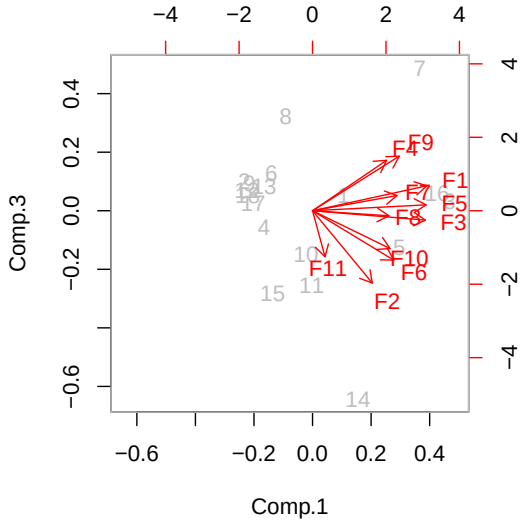
Per interpretare le prime componenti principali studiamo la matrice dei loadings e le loro proiezioni nei piani principali. Interpreteremo solo le prime 3 componenti principali, perchè le uniche con varianza maggiore di uno.

Per chiarezza dei grafici che useremo, abbiamo rinominato le regioni come segue:

- Da 1 a 7 rispettivamente: Piemonte, Liguria, Lombardia, Trentino Alto Adige, Veneto, Friuli-Venezia Giulia, Emilia Romagna.
- Da 8 a 16 rispettivamente: Abruzzo, Molise, Campania, Puglia, Basilicata, Calabria, Sicilia, Sardegna, Centro.
- 17 e 18 rispettivamente: Bolzano e Trento.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	Comp.7	Comp.8	Comp.9	Comp.10	Comp.11
F1	0.40	0.07	0.19	0.10	0.16	0.13	0.15	0.23	0.10	0.38	0.72
F2	0.21	−0.34	−0.54	0.21	0.34	−0.32	0.30	0.18	0.33	−0.22	−0.04
F3	0.40	−0.03	−0.07	0.16	0.06	−0.11	−0.48	0.31	−0.11	0.49	−0.46
F4	0.26	−0.34	0.38	−0.16	0.37	−0.36	−0.10	−0.60	−0.07	0.04	−0.10
F5	0.40	−0.05	0.04	0.10	0.08	0.19	0.12	0.17	−0.71	−0.48	−0.02
F6	0.29	−0.28	−0.37	0.03	−0.17	0.64	−0.23	−0.42	0.17	0.16	0.05
F7	0.30	0.45	0.11	0.20	0.13	0.14	0.58	−0.29	0.15	0.14	−0.41
F8	0.27	0.50	−0.04	0.25	−0.18	−0.32	−0.42	−0.17	0.22	−0.43	0.21
F9	0.30	−0.15	0.41	−0.45	−0.14	0.13	0.13	0.36	0.46	−0.31	−0.20
F10	0.27	−0.05	−0.28	−0.41	−0.62	−0.38	0.23	−0.07	−0.20	0.19	0.10
F11	0.04	0.45	−0.35	−0.64	0.50	0.10	−0.09	0.03	−0.06	0.03	0.01





Possiamo dedurne alcune considerazioni:

- La prima componente dipende positivamente da F1, F3 e F5
- La seconda componente dipende positivamente da F7, F8, F11
- La terza componente dipende negativamente da F2, F6 e positivamente da F9

Tuttavia, dalle tabelle e dalle osservazioni fatte finora non è immediato assegnare F4 e F10 (infatti dai loadings si vede come più componenti hanno valori comparabili relativi a questi fattori).

Un modo per risolvere questo problema è "cambiare punto di vista" con una rotazione opportuna delle prime tre componenti principali e di seconda e terza componente principale tramite la funzione *varimax*.

```
##
## Loadings:
##     Comp.1 Comp.2 Comp.3
## F1   0.352  0.271
## F2  -0.110 -0.169 -0.643
## F3   0.198  0.216 -0.276
## F4   0.551 -0.160
## F5   0.300  0.167 -0.215
## F6                -0.538
## F7          0.537
## F8          0.569
## F9   0.526
## F10         0.104 -0.382
## F11 -0.377  0.409 -0.112
##
##                Comp.1 Comp.2 Comp.3
## SS loadings     1.000  1.000  1.000
## Proportion Var  0.091  0.091  0.091
## Cumulative Var  0.091  0.182  0.273
```

```
##
## Loadings:
##     Comp.2 Comp.3
## F1          0.198
## F2         -0.640
## F3
## F4  -0.488  0.152
## F5
## F6         -0.462
## F7   0.329  0.327
## F8   0.452  0.223
## F9  -0.337  0.277
## F10        -0.270
## F11  0.562
##
##                Comp.2 Comp.3
## SS loadings     1.000  1.000
## Proportion Var  0.091  0.091
## Cumulative Var  0.091  0.182
```

A questo punto, guardando il risultato di varimax, assegniamo F4 alla prima componente e F10 alla terza componente. Finalmente, siamo in grado di interpretare le prime tre componenti principali:

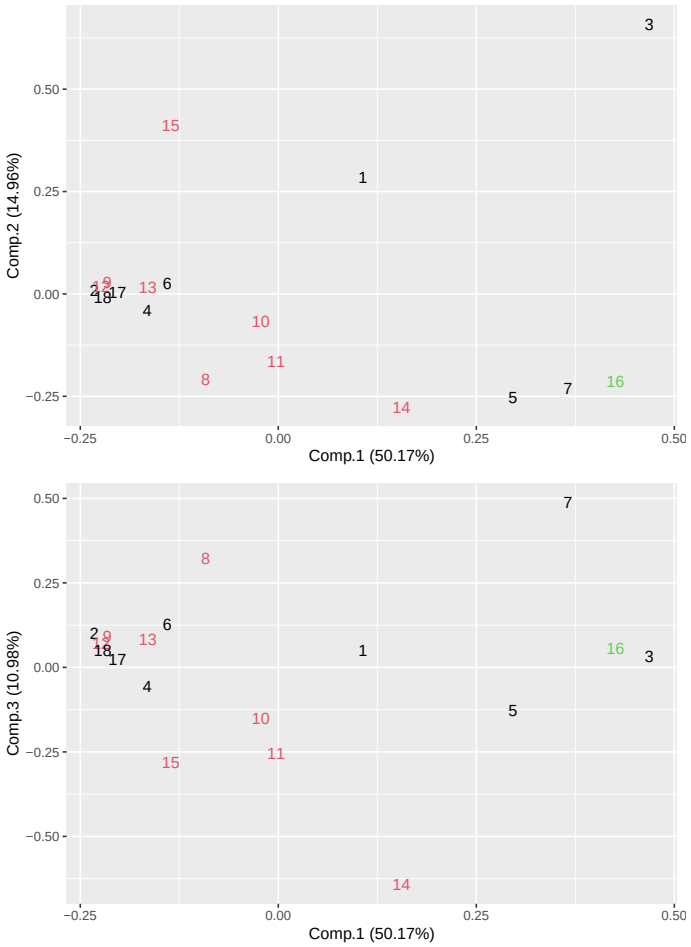
- La prima componente dipende positivamente da F1, F3, F4 e F5, quindi possiamo interpretarla come l'**utilizzo totale dei concimi**
- La seconda componente dipende positivamente da F7, F8, F11, quindi possiamo interpretarla come la **produzione di mais, loietto e altri erbai monofiti**
- La terza componente dipende negativamente da F2, F6, F10 e positivamente da F9, quindi possiamo interpretarla come la **produzione di erba medica e prati permanenti in funzione di concimi minerali a base di meso-elementi e micro-elementi e di prodotti fitosanitari**. Precisamente, la produzione di erba medica è inversamente proporzionale rispetto all'uso di concimi minerali a base di miso e micro elementi e di prodotti fitosanitari, mentre la produzione di prati permanenti è direttamente proporzionale a questi fertilizzanti.

Forti delle intepretazioni delle componenti appena date, commentiamo la proiezione delle componenti sul

piano principale (quindi rispetto al piano generato dalle prime due componenti): osserviamo che le prime 2 componenti riescono a distinguere la maggior parte delle regioni italiane, anche se è presente un insieme di regioni che si addensa nella parte sinistra del piano. Tuttavia, la proiezione rispetto alla seconda ed alla terza componente principale riesce (parzialmente) a "scioglierlo".

Inoltre, notiamo come Sardegna, Piemonte e Lombardia (ma per quest'ultima serve un discorso a parte, ci torneremo) si collocano nella parte alta del piano principale, deduciamo (dall'interpretazione delle componenti) quindi che queste regioni presentano vaste superfici di colture di mais, loietto e altri erbai monofiti, utilizzando una gran quantità di concimi. Discorso analogo e opposto si può fare per le regioni che si collocano nella parte bassa del grafico.

Indichiamo con colori differenti regioni di provenienza differente (nero per il Nord, verde per il Centro e rosso per il Sud) e rappresentiamole nei piani principali (i.e. i piani con componenti rispettivamente 1,2 e 1,3).



Notiamo che la Lombardia ha un comportamento molto particolare nel primo piano principale rispetto alle altre regioni: ci aspettiamo che sia un outlier. Per verificare o smentire il claim, proviamo ad indagare sulla stabilità dell'analisi: per ogni i osservazione valutiamo le componenti principali dei dati senza i e vediamo quanto si discosta la predizione di i dal risultato che ci dà il modello di partenza.

Riportiamo per semplicità i residui ad ogni passo:

```
##  [1] 0.043304523 0.006626157 7.172469567 0.001897257 0.016825321 0.002201678
##  [7] 0.166375935 0.066414620 0.005452063 0.002432732 0.033498472 0.006644549
## [13] 0.003579191 0.370314084 0.660718001 0.026109821 0.004178700 0.006898574
```

Effettivamente, togliendo la terza osservazione (i.e. la Lombardia) si ottiene un residuo non trascurabile,

a differenza di tutti gli altri casi. Possiamo concludere che la nostra intuizione è vera, in altri termini la Lombardia è un outlier.

Guardiamo ora come si collocano le regioni sulla proiezione rispetto alla prima e alla terza componente: per come abbiamo interpretato le componenti, questo grafico ci aiuta a capire la produzione di erba medica e di prati permanenti in funzione di alcuni fertilizzanti. Notiamo che Emilia Romagna e Abruzzo spiccano su tutte e che anche in questo grafico non riusciamo a distinguere bene Nord e Sud.

Conclusione

L'analisi delle componenti principali riesce certamente a semplificare il problema: difatti, su 11 fattori iniziali già le prime 3 componenti principali riescono a catturare il 76% di varianza spiegata. Abbiamo inoltre interpretato le prime 3 componenti principali, riuscendo anche a studiare le informazioni date dalle osservazioni (i.e. dalle regioni).

Precisamente, dallo studio del primo piano principale risulta che Sardegna, Piemonte e Lombardia presentano un'ampia area dedicata alle colture di mais e loietto, tutt'altra storia invece per le ragioni del sud. Dallo studio del secondo piano principale (i.e. quello relativo rispetto a prima e terza componente) risulta che Emilia Romagna e Abruzzo presentano un'ampia area dedicata alle colture di erba medica.

Ne deduciamo, infine, che le colture foraggere e l'uso dei fertilizzanti sono strettamente collegati (come si vede nell'interpretazione delle componenti), sia per la produzione di mais e loietto che per la produzione di erba medica.

Dataset utilizzati

Il dataset utilizzato è frutto di alcune tabelle presenti nel sito dell'Istat http://dati.istat.it/. Precisamente, si trovano nelle sezioni:

- Agricoltura → Mezzi di Produzione → Fertilizzanti → Fertilizzanti distribuiti per stato liquido o solido
- Agricoltura → Mezzi di Produzione → Fitosanitari → Prodotti distribuiti
- Agricoltura → Coltivazioni e allevamenti → Foraggere