

R と階層型クラスタリングを用いた主要指標の分析

工学部情報システム工学科 3 年 名列番号 065 山本祥悟

1. R について

R とは統計解析ソフトウェアのひとつである。

オープンソースでフリーソフトであることから、世界中の研究者がパッケージと呼ばれる拡張機能を公開しており、利用者はこれらのパッケージの中から自分の利用したいものを選んで機能を拡張できるのが特徴である。

R は下の URL からダウンロード可能。

<http://www.r-project.org/>

2. 学習方法

舟尾暢男氏がインターネット上で公開しているテキスト R-Tips をテキストとして使い、実際に R を動かしながら学習を行う。

R-Tips は下の URL からダウンロード可能。

<http://cse.naro.affrc.go.jp/takezawa/r-tips.pdf>

3. クラスタリング(クラスター分析)とは

クラスタリングとは、まだ分類されていないいくつかの事象を、似たようなもの同士にグループ分けすることを言う。クラスターとは英語で「集団」を表し、一見相互関連がなさそうなデータ集合の中から、似たような値、変化を表している事象を取り出すのが目的である。

4. 解析内容・方法

今回の自主課題研究では石川県のホームページで公開されている主要指標データを利用した。

<http://toukei.pref.ishikawa.jp/dl/1434/qis106.xls>

このエクセルファイルには、

- ・ 平成 14 年から平成 18 年までの 5 年間の年別変化データ
- ・ 平成 18 年 10 月から平成 19 年 9 月までの 12 ヶ月の月別変化データ

が全国・石川県で用意されている。これら全 4 種類のデータに対し、R によるクラスタリングを行うことで似たような変化をしている指標データをまとめる。なお、以下では全国の月別変化のデータに絞って説明する。

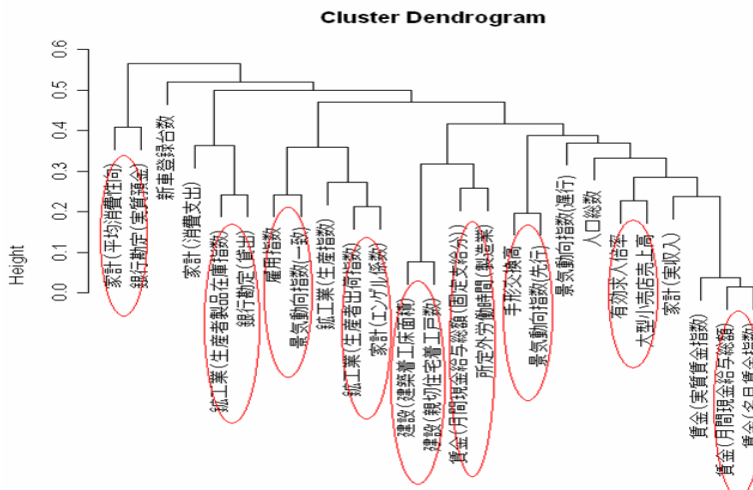


図 1 主要指標のクラスタリング結果
(全国、平成 18 年 10 月～平成 19 年 9 月)

その後、クラスタリングによってまとめられた各指標が実際にどのような変化をしているかを確認し、クラスタリングが有効であったのかどうかを確認する。

5. 解析結果・考察

クラスタリングの結果が下の図 1 である(出力に赤丸はなく、あとで私が記入した)。このように、クラスタリングの結果は樹形図(デンドログラム)として表される。これは近い部分にあるものが似た特徴を持つことを示すため、今回の樹形図では赤丸を記入した部分が似たような変化の特徴を持っている、ということを示している。

この図では似たような特徴を持っていると解析されたが、その関連性がよくわからない例もある(例えば今回では手形交換高と景気動向指数(先行)の変化が連動するのは何となく分かるが、鉱工業(生産者製品在庫指数)と銀行勘定(貸出)が連動している理由については疑問が残る)。

そこで指標の変化の様子を実際にグラフとして表し、クラスタリングの信頼性を確かめる。

先ほど疑問が残ると記した、鉱工業(生産者製品在庫指数)と銀行勘定(貸出)の実際の変化の様子をグラフにしたのが下の図 2 である(左が鉱工業、右が銀行勘定)。

縮尺は違うが、同じような変化をしていることが読み取れる。

6. まとめ・感想

特徴のわかりにくいデータの羅列から、クラスタリングによって似たような変化を表している事象を抽出することができた。

今後の課題として、抽出された事象が「なぜ」類似性を示すのか、その理由について考えることが必要である。それに加えて年別変化データと月別変化データによる比較、全国と各都道府県データによる比較などができれば、より考察に幅が出たと思う。もう少し早く課題設定を行うことができればこのような問題にも手を付けることができたと思うので、もう少しスケジュールに気を使う事ができればよかった。

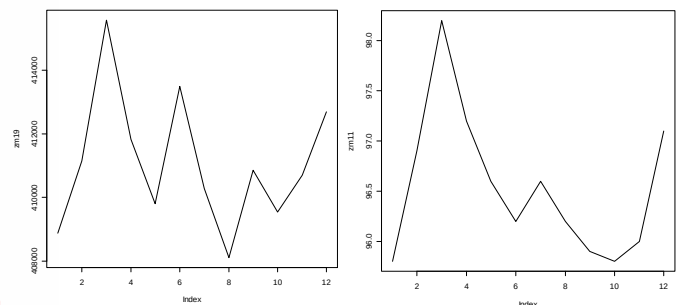


図 2 鉱工業(生産者製品在庫指数)(左)と
銀行勘定(貸出)(右)の実際の変化の様子