

共同研究「経営実務データを用いたデータサイエンス
ティスト育成方法の研究」報告書

2022 年 3 月 31 日

国立大学法人滋賀大学データサイエンス学部長

竹村 彰通

1. 共同研究の趣旨

本共同研究の研究題目は「経営実務データを用いたデータサイエンティスト育成方法の研究」であり、本共同研究の目的は企業の経営実務データを用いて販売予測や適正な在庫管理など、データサイエンティストに求められる能力を育成するための方法を継続的に研究することである。

特定非営利活動法人ビュー・コミュニケーションズは実際の販売データを教育目的のために適切に処理をおこなった上で国立大学法人滋賀大学に提供し、滋賀大学はデータをデータサイエンス学部の2年生対象の「時系列解析入門」の演習の形に整備して学生の教育に活用することで、データサイエンティスト育成方法に関する共同研究をおこなった。本年度は昨年度に引き続きビュー・コミュニケーションズから、食品スーパーにおける5年分の酒類の月次販売データが提供された。このような実際のデータを用いた演習は本共同研究の目的に沿うものであり、実践的なデータサイエンティスト育成の観点からも非常に有用であった。その点で本共同研究の意義は大きい。

2. 滋賀大学データサイエンス学部及び大学院の教育方針と実績

滋賀大学データサイエンス学部では、データサイエンスの基礎要素技術としてのデータエンジニアリング(情報工学)及びデータアナリシス(統計学)に加えて、データからの価値創造の能力の育成を重視している。価値創造の能力を育成するためには、実際のデータを用いた演習をおこない、教科書的な理論と実際のデータとの乖離についても経験を積む必要がある。この乖離については、ビュー・コミュニケーションズ副理事長の小松秀樹氏の著書「なぜあなたの予測は外れるのかーAIが起こすデータサイエンス革命」においても多くの例とともに示されており、データサイエンス教育において実際のデータを扱うことが非常に重要であることがわかる。

滋賀大学データサイエンス学部は、2017年4月に日本初のデータサイエンス学部として開設以来、この3月で5年となり2期生が卒業となった。卒業生の就職状況は大変よく、データサイエンス分野に対する社会のニーズの強さを示している。また1期生2期生とも、1学年定員100名のうち2割程度の学生が修士課程に進学した。大学院への2割の進学率は、国立大学における文系と理系の大学院への進学率の中間的な値であり、データサイエンス学部が文理融合的な性格を持っていることを反映している。

滋賀大学では2019年4月には、日本初のデータサイエンス研究科修士課程を開設し、昨年3月には学部の1期生と同時に修士課程の1期生も卒業となった。修士課程の開設は、データサイエンス学部の開設から2年後のことであり、まだ学部からの卒業生が出ていない段階での「前倒し設置」であった。学部からの進学性がないため、入学者のほとんどは企業からの派遣の院生であった。その後も毎年15名程度の派遣社会人が修士課程に入学している。これらの社会人は、それぞれ派遣元の企業の実際の課題を解決するための修士論文研究をおこなっている。

また 2020 年 4 月には博士後期課程も開設した。このように滋賀大学では学部から博士課程まで一貫したデータサイエンス教育体制を完成させ、日本におけるデータサイエンス教育研究の拠点としての体制を整えた。

3. 滋賀大学データサイエンス学部の時系列解析入門講義の概要

ここでは本共同研究で提供されたデータを活用した時系列解析入門の講義の概要を説明する。この講義はデータサイエンス学部の 2 年生の秋学期におこなわれる講義で、学年の 6 割以上の学生が受講する。教科書は以下の英語の教科書を用いた。

Cowpertwait and Metcalfe 著 “Introductory Time Series with R”, Fourth Edition, Springer Use R! series, 2009 年

英語の教科書を用いたのは以下のような理由による。まず、時系列解析については、日本語では応用に重点をおいた教科書が少ない。その点、この Cowpertwait and Metcalfe の本には多くの実例が示されており、使われているデータファイルも提供されている。次に、データサイエンスや AI に関する最新の文献は英語のものが圧倒的に多く、データサイエンティストは英語の文献を読むことも要求される。最後にこの教科書は電子書籍として滋賀大学の図書館より入手することができるため、学生が無料で利用できる。なお、ほとんどの学生は英語の教科書で学ぶのが初めてであり、その意味では難しいと感じた学生も多かったことも事実である。

さて、この教科書の目次は以下のようにになっている。

1. Time Series Data (時系列データ)
2. Correlation (相関)
3. Forecasting Strategies (いくつかの予測法)
4. Basic Stochastic Models (基礎的な確率モデル)
5. Regression (回帰分析)
6. Stationary Models (定常モデル)
7. Non-stationary Models (非定常モデル)
8. Long-Memory Processes (長期記憶過程)
9. Spectral Analysis (周波数解析)
10. System Identification (システム同定)
11. Multivariate Models (多変量時系列)
12. State Space Models (状態空間モデル)

講義では、8, 10, 11 章は入門としては進んだ話題のため省略した。講義の第一の目的は時系列解析の標準的なモデルである ARIMA モデルを理解し使えるようになることであるが、この教科書では ARIMA モデルは 6 章及び 7 章で扱われている。6 章は定常モデルとあるように ARMA モデルを扱っており、7 章では差分操作を含む ARIMA モデルを扱っている。また季節差分の扱い(SARIMA モデル)も 7 章で扱われている。本共同研究のデータを用いたレボ

ート課題は7章を終わった時点で課題として出題し、SARIMA モデルを中心に分析をするようにという指示を与えた。

4. 本年度ビュー・コミュニケーションズより提供されたデータ

本年度にビュー・コミュニケーションズより提供されたデータは、昨年度と同様に関東圏食品スーパー100 店舗サンプルの販売金額合計値時系列データであった。販売品目は酒類10 品目であった。データはエクセルのファイルとして、すでに整形された形で提供されており、学生はデータを R や Python などの統計分析パッケージに読み込むことで、ARIMA モデルなどの時系列解析の手法を応用することができた。このデータは時系列解析の例題としては非常に実践的なものであった。次図は提供されたエクセルファイルの一部を示したものである。

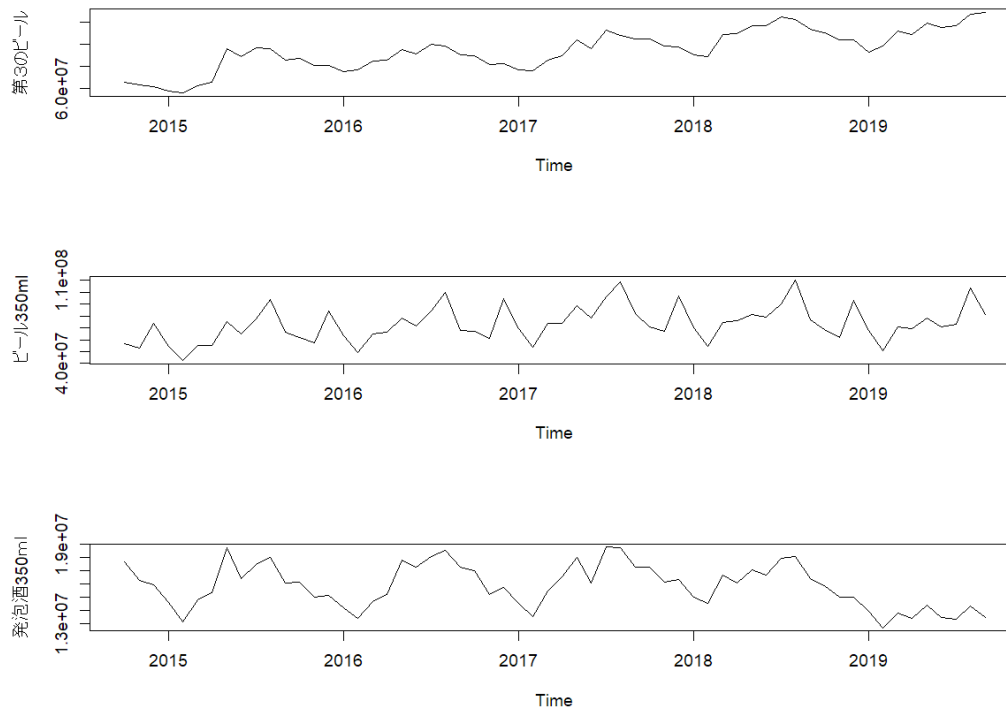
	第3のビール	ビール 350ml	焼酎甲類	清酒パック	国産ワイン	輸入ワイン	国産ウイスキー
2014年10月	64,896,265	56,458,480	41,283,034	27,710,083	15,671,871	17,012,474	11,630,853
2014年11月	62,626,576	53,164,431	42,021,722	30,565,834	17,547,875	33,758,505	12,350,601
2014年12月	61,599,319	73,810,979	44,625,092	37,739,137	20,952,974	25,209,042	14,292,028
2015年1月	56,903,102	54,636,556	39,416,436	28,568,680	15,394,759	18,412,583	13,251,262
2015年2月	55,275,135	42,429,178	37,706,181	28,192,179	14,140,309	16,260,936	13,442,114
2015年3月	61,832,614	55,360,190	41,023,819	28,121,201	15,395,572	16,849,718	15,683,896
2015年4月	65,137,971	55,137,113	39,079,248	25,829,663	13,980,098	16,657,483	15,533,525
2015年5月	95,580,312	75,617,063	43,269,560	22,720,433	13,507,734	17,591,824	14,537,782
2015年6月	88,410,801	64,505,572	55,492,424	20,548,710	11,904,098	15,005,699	13,854,597
2015年7月	96,231,198	77,867,879	39,422,324	19,173,130	11,388,621	14,482,564	13,158,490
2015年8月	95,740,071	93,823,473	38,529,894	18,322,581	10,929,671	14,653,894	13,122,023
2015年9月	85,365,554	66,422,417	38,211,951	22,253,526	12,521,920	15,458,816	12,324,527
2015年10月	87,278,610	61,513,467	40,727,491	27,928,000	16,921,779	19,699,321	13,375,270
2015年11月	80,622,982	57,462,902	41,008,635	31,193,441	17,679,276	34,439,141	13,335,337
2015年12月	80,550,355	84,451,103	43,382,753	37,712,380	20,849,978	26,678,527	15,593,571
2016年1月	74,896,265	63,204,148	39,193,684	30,669,160	15,611,565	19,267,511	13,578,948
2016年2月	76,489,294	49,531,205	37,564,149	29,950,650	14,643,939	18,477,289	13,249,479
2016年3月	84,751,597	64,607,404	40,332,548	30,191,077	14,764,217	19,224,918	14,284,018
2016年4月	85,530,588	66,563,713	38,443,524	27,844,328	14,134,252	19,286,489	15,331,135
2016年5月	94,723,466	78,050,551	42,703,326	25,602,536	14,566,449	21,522,904	16,997,167
2016年6月	90,770,895	71,194,568	51,544,702	22,761,122	12,312,624	17,275,704	18,719,842
2016年7月	99,618,096	84,006,041	39,677,949	21,045,965	11,559,434	16,372,270	19,395,526
2016年8月	97,794,657	99,688,086	38,734,847	19,953,496	10,942,939	15,548,213	19,338,026
2016年9月	89,843,497	67,595,129	37,984,195	22,357,293	12,079,329	16,311,333	19,512,245

本年度提供データの一部

5. 学生によるデータ分析の例

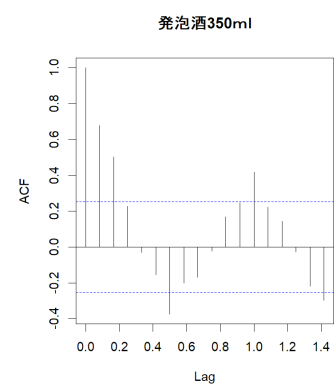
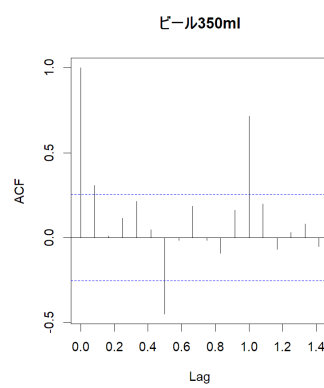
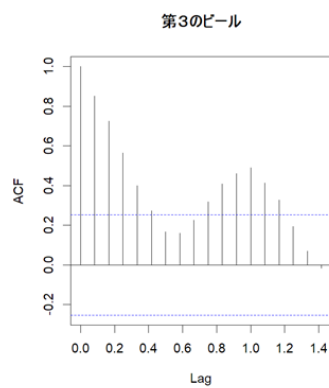
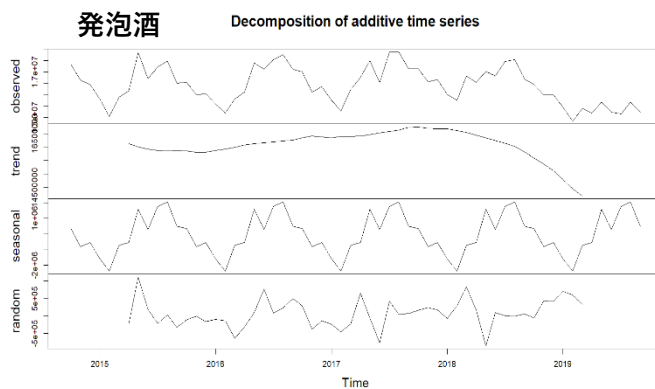
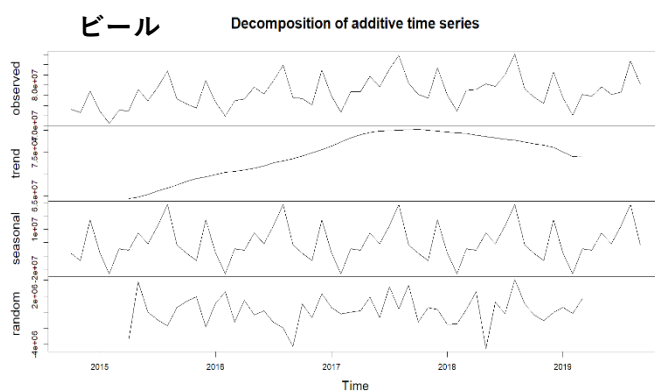
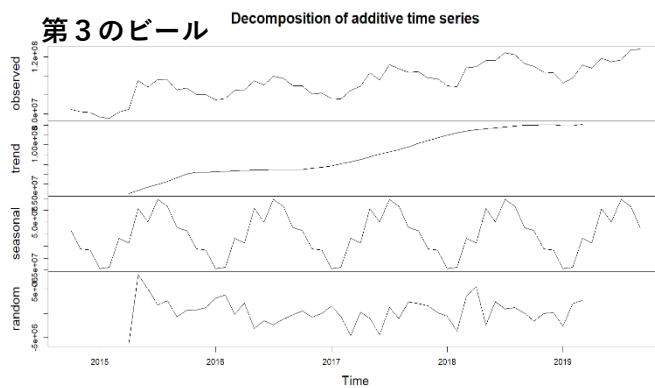
レポート課題では、ビュー・コミュニケーションズより提供された 10 品目のデータの中から 3 品目を選んで解析するよう学生に求めた。以下は学生の川口陽子さんによる分析である。

第3のビール、ビール、発泡酒の今後の売上を予測する



第3のビール、ビール、発泡酒の売り上げに関してプロットしたものである。1年ごとに周期的に推移していることがわかる。また、第3のビール、ビールは右上がりのトレンドを示しているのに対して、発泡酒は2018年から減少傾向を示しているように見える。

よりトレンドを見やすくするためにトレンドと季節成分と誤差成分に分解して表示した



ものが以下である。第3のビールについては、上昇するトレンドが続いていることがわかる。

一方で、ビールと発泡酒に関しては、第3のビールの上昇傾向が強くなった2017年中ごろから2018にかけて下降傾向に転じていることがわかる。

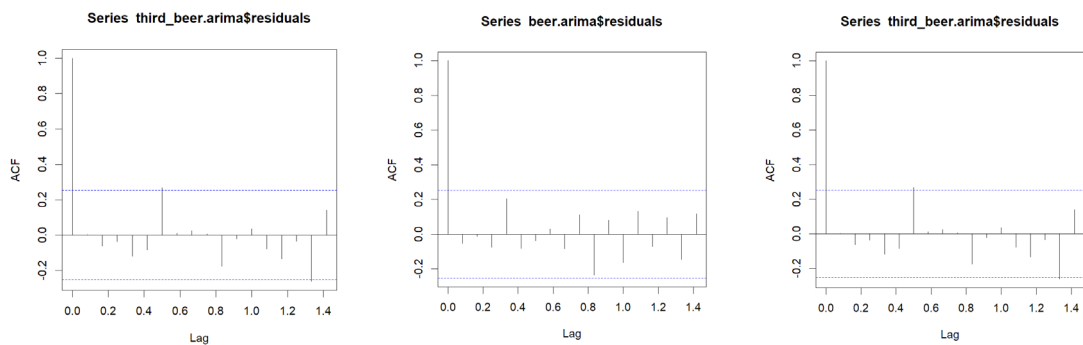
近年、ビールや発泡酒に比べて酒税がかからず安価に手に入る第3のビールが普及したことでビールや発泡酒の売上が減少している可能性がある。

そこで、今後、ビールや発泡酒が第3のビールに置き換わるのかを予測していきたい。まず、それぞれの過程の自己相関に着目する。それぞれのコレログラムを見ると、すべてにおいて12か月ごとに相関が高くなっており、明らかに季節成分があることが確認できる。

よって、季節性についても考慮されたSARIMAモデルを用いて予測を行う。

#条件を以下の様にする。

```
third_beer.arima <- auto.arima(third_beer.ts,  
                                d=1,  
                                ic="aic",  
                                max.order=5,  
                                stepwise=F,  
                                seasonal=T,  
                                trace=T  
                                )
```

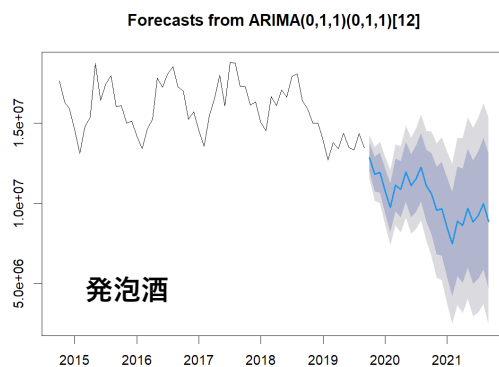
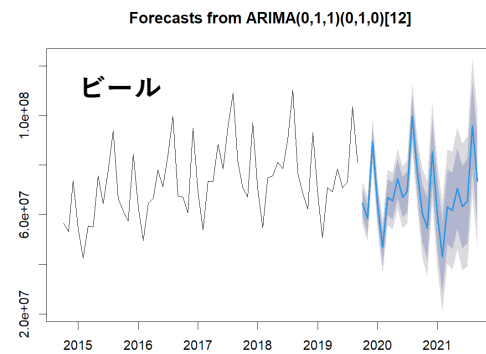
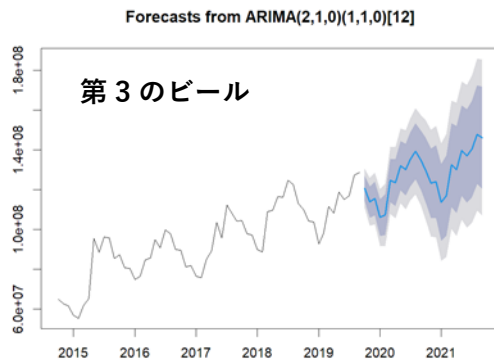


3つの時系列過程について、1回差分を取ったところ、その過程が発散しておらず、かつ Dicky-Fuller 検定が有意水準5%で棄却されたことから定常過程になっていることが確認できた。そこで、パラメータ d について1とし、その他の条件を上図のように設定して SARIMA モデルの推定を行った。ここでは、例として第3のビールで用いたコードを記載している。

その後、モデルが適しているかどうかを確認するために、残差のコレログラムを描いて自己相関がないかを調べたところ、十分に小さくなっていた。

次ページに、これらのモデルから2年後までの予測値をプロットしたものが記載されている。第3のビールについては、2020年以降も上昇する傾向が見られている。一方で、ビールについては、売上が横ばいかやや下降傾向になることが予測されている。また、発泡酒についても急激な下降傾向に歯止めがかからない予測となった。

このことから、安価な第3のビールが今以上に普及しても、ビールについてはある程度の需要を保つことができるが、発泡酒については徐々に第3のビールに置き換わられていくのではないかと考えられる。



以上のように川口さんは、データのプロットから始め、情報量規準を用いて最適な SARIMA モデルを選び、また選ばれたモデルの当てはまりまでを検討しており、標準的な分析例となっている。このような分析ができる学生がいたことは、この講義の成果であると考えられる。

6. まとめ

前節でみたように、提供されたデータの分析のレポート課題に多くの学生が熱心に取り組んだ。その理由は、提供されたデータが非常に実践的で、また時系列解析の効果が実感できるものであったことである。

滋賀大学データサイエンス学部としては、ビュー・コミュニケーションズとの共同研究のような形で多くの企業との連携を進め、実践的な教材を今後も開発していきたいと考えている。滋賀大学は数理及びデータサイエンスに係る教育強化の拠点校として文部科学省より選定されており、このような教育コンテンツを他大学に展開していくことが期待されている。日本初のデータサイエンス学部及び研究科として、この分野の発展に貢献していきたい。