

ストリームデータの動的統計分析 —統計処理教育へのアプローチ—

吉 岡 茂*

キーワード：ストリームデータ、スライディングウィンドウ、ビッグデータ、IOT、VBA

1 はじめに

ストリームデータは気象情報や為替・株価などの金融情報、医療機関でのバイタルデータ、交通情報、ツイートデータのような時刻順に到来する終わりのないビッグデータのことで川の流れに例えられる。近年の産業や社会におけるIOT（Internet Of Things）の普及とデータサイエンスの進歩によって、これまで気づかずに捨てられてきたビッグデータや、気づいていてもデータ取得技術の未発達により見過ごされてきたデータを処理することが可能になった。

IOTで得られた交通情報を利用すれば信号機の切り替え時間の最適化が可能となり、さらに気象情報を組み合わせることで高い精度で交通事故の発生確率を予測し、運転者に対して事故発生可能性を警告することが可能になる。また目的地までの到着時間の推定の精度を高めることも可能となり、現在のカーナビ機能の大幅な向上による運転者の安全確保と負担軽減に貢献でき、省エネルギー対策としても有効である。

表1 ストリームデータと利用例

データ	利用例
・金融情報（為替、株価）	損失回避
・犯罪発生予測	犯罪防止
・交通情報	事故予防・渋滞回避
・バイタル	救急治療
・消費電力	ブラックアウト回避
・気象情報	熱中症予防
・放射線情報	被ばく予防

ストリームデータ処理技術は、データベースに蓄積された過去のデータを扱う「ストック型」の伝統的なデータ処理技術とは異なる「動いている終わりのないデータ」を扱う新しい技術である。通常のデータ処理は固定

されている過去のデータを一括して分析するが、ストリームデータ処理では、データが発生する都度継続的に処理を行う。そのためストリームデータ処理結果は刻々と変化し、一般的には古いデータは徐々にその重要さを失ってゆくことが多い。またデータの変化に応じて処理方法（分析方法）をダイナミックに変更することもできる。

本論は情報処理や統計処理教育を主眼に、時間軸に沿ってダイナミックに変化する東京証券取引市場や上海市場等の株式や為替レートに関するストリームデータを利用した動的統計分析を試みたものである。

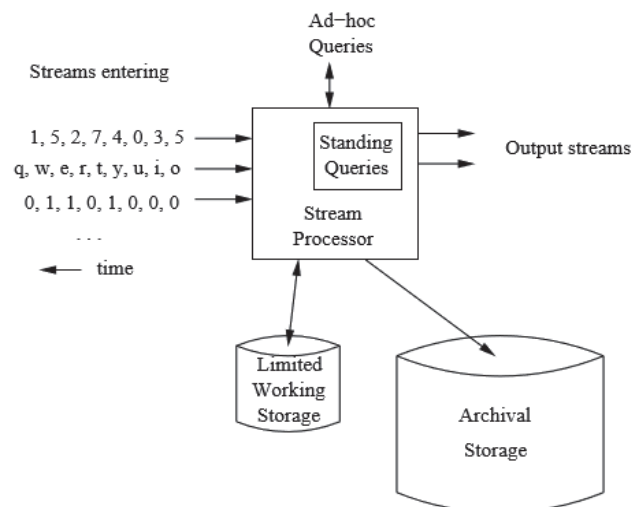


図1 A data-stream-management system

(^[1] Jure Leskovec, Anand Rajaraman, Jeff Ullman)

2 ストリームデータの動的処理

無限に流れ込んでくるストリームデータの中から標本を選ぶ方法のひとつにスライディング（Sliding）ウィンドウがある。例えば1秒ごとに到来する時系列データの処理を100秒間に到達した100標本について処理し、次の処理は1秒後のデータを最新データとしてすでに到達済

* 立正大学地球環境科学部

みの99標本と合わせた100標本について再度処理する方法である。実際の統計処理では、一番古い標本1個を減らし、新たに到達した標本1個を付け加える差分処理を行うことになる。

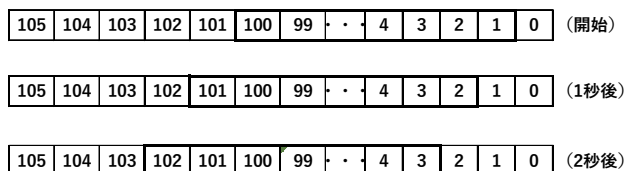


図2 スライディングウィンドウの概念
〔² 日立製作所〕

2.1 ストリームデータ

本論では極めて短時間に大量のデータが発生し、しかもインターネットを使用することにより無償でデータ入手の可能な東京証券取引所の株価と為替データを利用した動的統計処理を行う。2018年12月1日現在の東京証券取引所で扱う上場企業数は全体で3,635、そのうち第一部上場企業2,111社、第2部504社、新興企業市場992社、プロ市場28社である。一般に東京市場の株価は、直前の株価、ニューヨーク（NY）証券取引所のダウ平均株価、基軸通貨ドルに対する円の為替レート及び重大経済ニュース等によって決まると考えられる。株価指数Nikkei225は東京証券取引所市場第一部の225銘柄の株価から作成される指数であるため、225銘柄それぞれの株価の変動が指数に影響を及ぼす。世界の主要証券市場の取引時間（日本時間）は休憩時間を挟んで次のようになる。

ニューヨーク（NY）市場 23：30～6：00
（サマータイム 22：30～5：00）

東京市場 9：00～15：00

上海市場 10：30～16：00

ロンドン市場 16：00～0：30

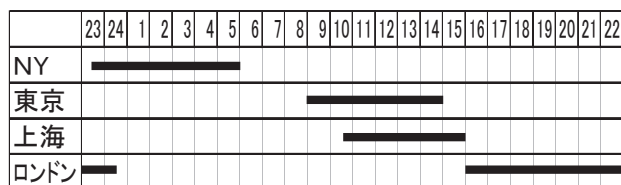


図3 主要証券取引所の取引時間（日本時間）

2017年の世界の主要5大証券取引所の取扱高を示したのが表2である。ニューヨーク市場（NY）の取扱高は5大市場の48.6%を占め、さらに米国のNASDAQ市場の19.5%と合わせると68%に達し、金融市場におけ

る米国の影響力の大きさは圧倒的である。NASDAQの株価指数はニューヨーク市場の株価指数Dowに代表されると考えると、東京証券取引所の株価指数Nikkei225は、世界の主要証券取引所の取引時間の関係からニューヨーク証券取引所の株価指数Dowの影響を受けて始まり、午前10時半からは上海市場の総合指数SSEの影響も受け始めると考えられる。ロンドン市場は東京市場閉鎖後に開くので、翌営業日にニューヨーク市場のDowを通じて間接的に東京市場に影響を及ぼすと考えられる。月曜日の東京市場のNikkei225は重大経済ニュースがない限り前週末のNikkei225とDow、取引開始時のドル円為替レート（DY）の影響を受けて始まり、その後上海市場のSSEの影響も受けるようになると考えられる。

表2 取扱高（2017）

取引所	取扱高（兆円）	割合（%）
NY	2,050	48.6
NASDAQ	820	19.5
東京	530	12.6
上海	425	10.1
ロンドン	390	9.3
計	4,215	100.0

2.2 モデル

以上から、Nikkei225を推計するモデルN225は次の様な関数で示される。

(1) $N225 = f(N225_{-1}, Dow, SSE, DY, News) + \varepsilon$
ここで、N225：Nikkei225、 $N225_{-1}$ ：1期前（直前）のNikkei225、Dow：ニューヨークダウ平均株価、SSE：上海総合指数、DY：ドル円為替レート、News：重大経済ニュース、 ε ：平均0、分散 σ^2 の正規分布に従う誤差項である。

東京市場の取引が始まる時点でニューヨーク市場は取引を終えているため、Dowは変数でなく定数となっているが、その増減率がNikkei225の大きな方向性に影響すると考えられる。そこで線形モデルを想定し、次のような重回帰式を考える。

$$(2) N225 = \beta_0 + \beta_1 N225_{-1} + \beta_2 DDow + \beta_3 SSE + \beta_4 DY + \beta_5 News + \varepsilon$$

ここで β_1 は偏回帰係数であり、DDowは前日に対するDowの増減率（ Dow/Dow_{-1} ）によって次のような値をとるダミー変数を仮定し、その値を取引開始時の $N225_{-1}$ に乗じて使用する。すでに終了しているニューヨーク市場のDowの上昇率または下落率が0.5%より大

きければ、前日の東京市場の終値 $N225_{-1}$ を0.5%それぞれ調整する考えである。従って、

(取引開始時)

$$DDow = \begin{cases} 1.005 & Dow/Dow_{-1} > 1.005 \\ 1 & 0.995 \leq Dow/Dow_{-1} \leq 1.005 \\ 0.995 & Dow/Dow_{-1} < 0.995 \end{cases}$$

(取引開始以降)

$$DDow = 1$$

となる。

モデル(2)ではニューヨーク市場の株価指数Dowの増減率の影響をN225に反映させるために、取引開始時の $N225_{-1}$ の初期値をDDowの大きさに調整する。Newsは重大ニュースがあった場合は1、なかった場合は0の値をとるダミー変数であるが、その重大ニュースが多くの投資家に知られるまでの間は1を取り、知られた後は0に変化する。しかし本論では単純化のためNewsをモデルから除外した。

3 データ取得と計算

3.1 VBAによるデータ取得

Nikkei225や為替レート等のストリームデータは、yahooのwebサイト(URL: <https://stocks.finance.yahoo.co.jp/stocks/detail/?code=>)からEXCEL VBA (Visual Basic for Applications) で読み込む。処理プログラムにEXCEL VBAを使用した理由は、インターネットサイトから直接データを読み込むことが出来るほか、EXCELと一体でデータ分析等処理が可能のためである。ただしこのサイトの為替レート等データは20分遅れのデータであるため、厳密には株価のようなリアルタイム情報とは同期していない。ここでは金融情報として上記yahooのwebサイトから、前日のダウ平均株価(Dow)、東京市場第一部上場企業の株価指数Nikkei225(N225)、上海総合株価指数(SSE)およびドル円(DY)の為替情報を取得している。重大ニュース(News)については、上述した通り本論では計算を簡略にするため説明変数から省略した。従って、最終的に使用したモデルは式(3)である。

$$(3) N225 = \beta_0 + \beta_1 N225_{-1} DDow + \beta_2 SSE + \beta_3 DY + \varepsilon$$

ここで12月1日(金)のDowの終値25,538.46、同3日(月)のDow終値25,826.43であるため、 $DDow = 1.005$ となる($25,826.43/25,538.46 = 1.0112 > 1.005$)。従って、12月4日の初期値 $Nikkei225_{-1}$ は前日の終値22,574.76に1.005を乗じて使用される。

ストリームデータの取得方法はwebサイトstocks.finance.yahoo.co.jpの中の株価データが記載されているソースコード(HTML)をVBAで読み込み、必要な株価指数等データを探して配列変数に蓄積したりEXCELシートに書き込む。また重回帰分析を行い、分析結果と利用者に対する注意喚起等の情報をMSGBOXでコンピュータ画面に表示する。



図4 stocks.finance.yahoo.co.jpの画面とソースコード(HTML)

例: Nikkei225のデータを取得するVBA

WEBサイトからVBA関数GetHtmlSourceを使用して、所要のデータを取得する。VBAでstocks.finance.yahooのソースコードからNikkei225の値を読み取る場合は、下記のstrURLに代入するhttps://の中でcode=998407.0を指定することで可能になる。上海総合指数(SSE)を取得する場合はcode=000001.SSを指定する。他の金融情報についても、同様の方法でそれぞれのcodeを指定することで取得可能になる。

(VBAプロシジャの一部)

```
strURL=https://stocks.finance.yahoo.co.jp/stocks/detail/?code=998407.0
strID = ""
strPass = ""
If Not GetHtmlSource(strURL, strRetVal, False, strID, strPass) Then
MsgBox "情報を取得不能", vbCritical
Exit Sub
End If
```

3.2 重回帰分析

こうして得たデータを利用して、重回帰式（2）を求める場合はEXCELの関数LINESTを利用したり、重回帰式を求めるVBAサブプロシジャーを作成して付加する。必要な統計量を求めさらにデータから自由な加工を行い見やすい形式で表示したい場合はVBAサブプロシジャーを特別に用意するのが良いが、ここではスライディングウインドウのサイズを $n=30$ として、EXCEL-VBAのLINEST関数を利用して偏回帰係数とその検定統計量 t 値、VBAプロシジャーにより自由度修正済み決定係数 (R^2) を求めて、コンピュータ画面とEXCELシートに順に一行ずつ出力した。なお、データは平均2秒につき新たに1標本を取得し、さらに一番古い1標本を除外した30標本について重回帰分析を行った。こうした処理を100回（30標本×100回）繰り返した。スライディングウインドウのサイズを $n=30$ とした理由は、ここで使用した金融データは1秒単位で激しく変動を繰り返す性格を有しているため、重回帰分析の自由度も考慮しながら、1分間程度の標本 $n=30$ を分析対象としたものである。一般にこのサイズを大きくすると実際のN225の変化に対する重回帰式から求めた予測値N225pの追従が鈍感になり、小さくすると鋭敏すぎて予測値が大きな傾向を見失うため、ウインドウサイズをいくつか変えてみることで妥当なサイズを見出す必要がある。

3.3 処理結果と考察

図5は2018年12月4日14時38分22秒から14時44分10秒までのスライディングウインドウのサイズを $n=30$ とし、標本を1個ずつずらして行った100回の重回帰分析による自由度修正済み決定係数と偏回帰係数等統計量の変化を示したものである。ドル円為替（DY）レートの変化に対し、重回帰式の定数項（ b_0 ）は逆方向に動くことを視覚的に確認できる。

処理時間が6分程度と短かったものの、この間に自由度修正済み決定係数 (R^2) は0.32から0.94まで大きく変化した。同様に表4から偏回帰係数の変化についてみると、定数項は最大値30,852、最小値-74,377、N225₋₁の偏回帰係数は最大値10,131、最小値4,100、上海総合指数（SSE）のそれが最大値1,050,674、最小値-38,103、ドル円レート（DY）のそれは最大値31,226、最小値-16,571であり、激しく変動していることが分かる。さらに各偏回帰係数の統計的有意性を示す t 値の変化を示したのが図6である。N225₋₁の t 値は常に大きな値を有し統計的に有意である。また全体として上海総合株価指数SSEの t 値は小さいが、5%有意水準2.06を上回る数値を示すことも観察された。ドル円為替レートDYも同様の挙動を示している。

金融データ自身が有する激しく変動する性格に起因するものの、わずか6分間程度のストリームデータから求めた重回帰分析の結果は大きく変化したが、N225

表3 偏回帰係数、 t 値及び自由度修正済み決定係数

NO.	自由度修正 済決定係数 R^2	偏回帰係数				t 値			
		const/10	N225 (₋₁) *100	SSE*100	DY*10	const	N225 (₋₁)	SSE	DY
1	0.8446	-3,231.0	88.0	-13.8	3,122.6	-2.192	9.749	-0.095	2.398
2	0.8519	-3,261.2	88.6	53.9	2,979.0	-2.146	9.431	0.392	2.234
3	0.8625	-2,910.0	93.4	17.7	2,660.8	-1.896	10.547	0.129	1.975
4	0.8883	-2,624.5	96.3	4.1	2,382.2	-1.811	12.513	0.032	1.873
5	0.8988	-2,614.1	96.5	7.3	2,362.0	-1.760	12.052	0.052	1.853
6	0.9062	-2,641.9	96.2	10.4	2,386.6	-1.778	12.746	0.072	1.885
7	0.9122	-2,580.9	96.8	2.2	2,339.5	-1.758	13.815	0.016	1.864
96	0.6574	-2,002.8	78.8	646.2	670.3	-0.868	6.776	0.705	1.195
97	0.7267	-2,576.7	68.4	1,195.2	94.4	-1.550	7.159	1.818	0.207
98	0.7787	-1,816.4	90.7	479.0	665.2	-1.236	8.587	0.795	1.428
99	0.7821	-1,650.2	92.0	395.9	687.7	-1.148	9.078	0.683	1.479
100	0.8118	-1,115.5	97.2	124.7	749.3	-0.838	9.421	0.227	1.677

の実際値とストリームデータから重回帰式で求めた予測値N225pを示したのが図7である。この図から予測値N225pは全体としては実際値N225をかなり高い精度で

追跡していることが分かる。N225に対するN225pの誤差率は0.0048%と極めて小さい。

$$\text{誤差率} : \sum \text{SQRT} \{ (N225p - N225) / N225 \}^2 / 100$$

表4 偏回帰係数、t 値及びR²の最大値最小値

	偏回帰係数		t 値		自由度修正済決定係数 (R ²)	
	最大値	最小値	最大値	最小値	最大値	最小値
B ₀	30,852	-74,377	2.834	-3.204	0.9436	0.3249
N225 ₋₁	10,131	4,100	19.120	2.347		
SSE	1,050,674	-38,103	3.196	-0.724		
DY	31,226	-16,571	2.398	-2.548		

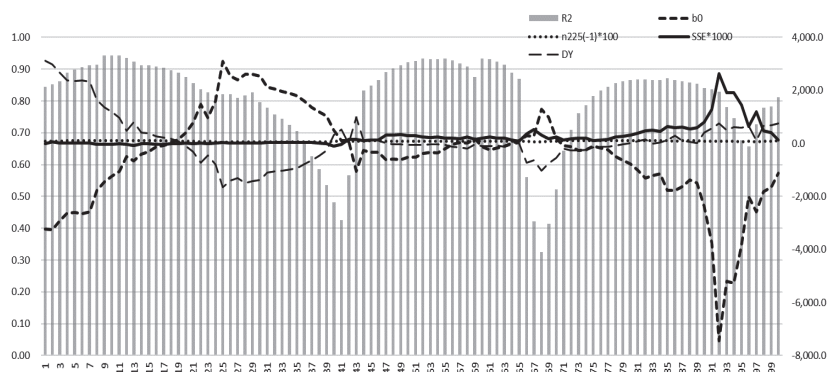


図5 統計量の変化

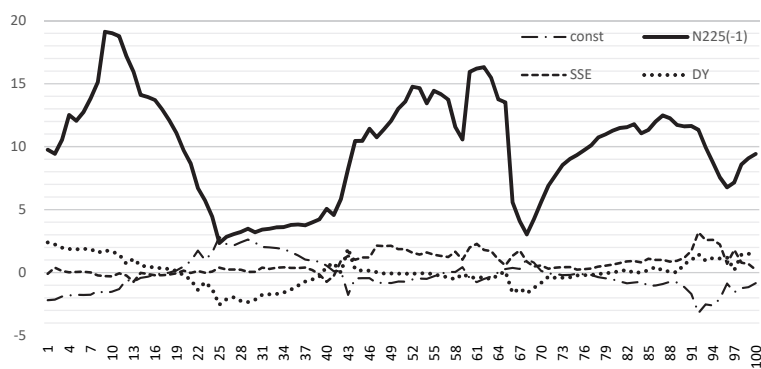


図6 各偏回帰係数の t 値

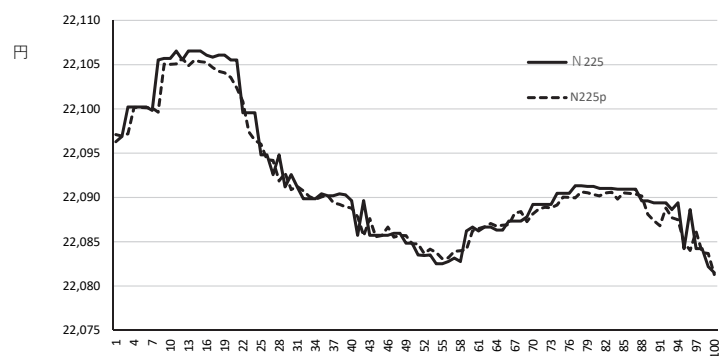


図7 N225と推定値N225pの関係

4 おわりに

本研究は情報処理教育を念頭に置いて開発したEXCEL-VBAによるWEBサイトからのストリームデータの取得と重回帰分析を例にとり、その統計処理について紹介したものである。同様の方法で、気象、電力、交通、通信量および放射線等に関するストリームデータの処理も可能である。ストックデータを利用した場合は、パターン発見や情報要約、クラスタリングなどの分析結果は一意に決定された固定的な構造が得られるが、ストリームデータを利用した場合は処理結果が時間とともにダイナミックに変化してゆくことになる。

ストリームデータに対する動的な統計処理法として、次のような方法論があげられる。

（1）変数選択による回帰分析

ストリームデータの中のある変数を回帰式で予測したい目的変数としたとき、説明変数の組み合わせの中からもっとも予測効率の高い組み合わせを選択し、偏回帰係数等統計量を出力するほか、得られた回帰式を使用して次期株価を予測する。スライディングウィンドウのサイズを指定することで、時間の推移に従って得られる回帰式も変化する。

（2）主成分分析

ストリームデータの中から関心のある複数のデータについて主成分分析を行い、固有値と固有ベクトルを求める。時間の推移とともに、第一主成分となる最大固有値の全固有値（計算対象とした変数の個数に一致する）に占める割合と固有ベクトルの符号と絶対値の変化を観測し、変数間の構造に変化が起きていないかどうかを調べる。さらに最小固有値が極めて小さい場合、固有ベクトルを係数とした変数間に線形関係がほぼ成立しているので、ある変数を他の変数から求めることが可能となる。株式データなら、ある銘柄の株価を他の銘柄の株価から推定できることになる。

（3）クラスター分析

ストリームデータの中の関心のあるデータの構造を視覚的にとらえたい場合に適している。株価のデータであれば、値上がり銘柄と値下がり銘柄がそれぞれどのような為替や海外株価指数の影響を受けているのかといったことをデンドログラムで表示できる。同じような動きをする銘柄をグルーピングし、デンドログラムで表示し、視覚でデータ構造の変化をとらえることが出来る。

（4）判別分析

株価のストリームデータだと、値上がり銘柄と値下がり銘柄に二分し判別式を求め、得られた判別式の係数から値上がり要因と値下がり要因を発見する。判別式を用いて値上がり銘柄と値下がり銘柄を予測し、判別効率（正しく判定した銘柄数/全銘柄数）が大きければ実用に耐えうる。また回帰分析と同様に、変数選択による判別式の導出も可能である。

今後の課題は、こういった方法論をストリームデータに適用することでその有用可能性を実証してゆくことである。また分析対象とするストリームデータのウィンドウサイズは、事前にいくつかのサイズで自動的に試行するVBAサブプロシジャを作成して、最適な大きさに決定することである。

参考文献

- [1] Jure Leskovec, Anand Rajaraman, Jeff Ullman, Mining of Massive Datasets, <http://infolab.stanford.edu/~ullman/mmds/ch4.pdf>, 2018
- [2] 日立製作所、「今」を分析するストリームデータ処理技術, <http://www.hitachi.co.jp/Prod/comp/soft1/cosminexus/sdp/download/HSDP.pdf>, 2018
- [3] SAS and all other SAS Institute Inc., Understanding Data Streams in IoT, https://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/understanding-data-streams-in-iot-107491.pdf, 2018

Dynamic statistical analysis of real-time streaming data

YOSHIOKA Shigeru*

* Faculty of Geo-environmental Science, Rissho University

Abstract:

In this article, We showed the example of the dynamic statistical analysis using a streaming data about the price index of stocks and the foreign exchange. Specifically, We read financial data from WEB Site in EXCEL-VBA and performed a multiple regression analysis to predict price index of stocks Nikkei225. We took Nikkei225 just before that, Shanghai stock prices Composite Index (SSE) and USDJPY in an explanation variable. This multiple regression analysis was repeated 100 times with the size of Sliding Window 30. Finally, We showed the characteristics of these statistics from the regression analysis.

Key words: Streaming Data, Sliding Window, Big Data, IOT, VBA

