3C3-3

状態空間モデルを用いた検索トレンドとページビューからの自動車販売台数の予測

Car Sales Prediction using State Space Model with Search Trend and Page View Data

角田 孝昭 *1 Takaaki Tsunoda　吉田 光男 *2 Mitsuo Yoshida　津川 翔 *1 Sho Tsugawa　山本 幹雄 *1 Mikio Yamamoto

*1筑波大学大学院 システム情報工学研究科
Graduate School of System and Information Engineering, University of Tsukuba

*2豊橋技術科学大学 情報・知能工学系
Department of Computer Science and Engineering, Toyohashi University of Technology

Search volume of search engines are expected to be effective for trend analysis as they reflect people's interest. In this paper, we propose several models for car sales prediction with search volume approximations. We used a search trend index (Google Trends) and a page view of the website which is ranked high in the search results (Wikipedia) as the approximations, and incorporated their trend component into our models based on a state space model. We evaluated our proposed models by predicting several car sales and results show that the proposed models outperform a baseline model without using search volume approximation.

1 はじめに

本研究では、購買行動に先立って行われる検索行動の動向に着目した、自動車販売台数の将来予測精度を改善するための手法について検討する。自動車は高額商品であることから多くの消費者が綿密な検索行動を行うため、検索行動量が販売台数に反映されると期待できる。検索行動の総量を推定するために、本研究では直接的な検索行動数を反映する Google Trends *1 に加え、実際の調査対象となるページへのアクセス数を反映する Wikipedia *2 ページ閲覧数を用いる。

自動車販売台数の推移は季節成分を伴った典型的な経済時系列であることから、予測モデルには状態空間モデルを用いる。本研究では、特に検索行動量の推移におけるトレンドに注目し、検索行動量トレンドを販売台数の予測に取り込むためのいくつかのモデルを提案する。また、実際に販売台数の予測を行う評価実験を通し、各モデルの有効性について議論する。

以下、2 章で検索行動量に基づく将来予測や状態空間モデルの応用に関する関連研究について述べる。次に、3 章で本研究で販売台数と検索行動量の相関などの性質について観察する。続く 4 章では、3 章の観察に基づいた予測モデルを提案する。5 章で評価実験を行う。6 章で本研究のまとめを行う。

2 関連研究

これまでに検索行動量として Google Trends を用いた研究は広く行われており、様々な予測において Google Trends を用いることで精度が改善することが示されている [Choi 12, Xu 12, Goel 10]。本研究の目的である自動車販売台数予測と最も関連が深い研究としては、Choi & Varian による自動車及び自動車部品ディーラー (Motor Vehicles and Parts Dealers) の売上金額を予測する研究がある [Choi 12]。これに対し、本研究では業界全体の総売上額ではなく、車種別の販売台数の予測と言うより細かい指標の予測を目的とする。なお、これに関連して Goel らは映画やゲームなどについて個別の商品ごとに売上予測を行っているが、彼らの対象はある固定された期間の合計売上数である [Goel 10]。加えて、以上の研究が Google Trends で得られた指数をそのまま説明変数として利用するのに対し、本研究では指数のトレンド成分を説明変数と考える拡張を行う。更に、以上の研究では独占データである検索トレンド指数に依拠しているが、本研究では Wikipedia の記事閲覧数と言うオープンなデータを用いた予測も新たに試みる。

また、状態空間モデルは時系列データの要因分解及び将来予測を可能にする強力な枠組みの一つであることから、これまでに様々な時系列への適用が試みられている。具体的な研究として、広告クリック率の予測 [本橋 12]、広告効果半減期の予測 [Naik 99]、通話料収入の予測 [矢田 93] などがある。本研究では状態空間モデルを自動車販売台数の予測へと適用すると同時に、検索行動量を活用することで予測精度の向上を試みる。

3 自動車販売台数と検索行動量との関係

3.1 データの入手

本研究で用いるデータは次のようにして入手した。なお、期間は 2010 年 1 月〜2015 年 2 月に固定している。

自動車販売台数については、日本自動車販売協会連合会 *3 (普通及び小型乗用車) 及び全国軽自動車協会連合会 *4 (軽自動車) が公開している毎月の新車販売台数データを利用した。このうち、2015 年 2 月時点でも発売されており、かつ上記の期間内で販売台数が比較的多い 22 車種を分析の対象とする *5。

また、検索行動量には Google Trends 及び Wikipedia 閲覧数を用いた。Google Trends については、車名にメーカー名を追加したキーワード (例えば「トヨタプリウス」など) に対する人気度値 (Interest) を利用した (以下、単に Google Trends 値と呼ぶ)。Wikipedia 閲覧数については、日本語 Wikipedia の該当する項目 (例えば「トヨタ・プリウス」など) に対する

連絡先: 角田孝昭. tsunoda@mibel.cs.tsukuba.ac.jp

∗1 Google Trends: http://www.google.co.jp/trends/

∗2 Wikipedia: http://ja.wikipedia.org/

∗3 日本自動車販売協会連合会: http://www.jada.or.jp/

∗4 全国軽自動車協会連合会: http://www.zenkeijikyo.or.jp/

∗5 プリウス, カローラ, パッソ, ノア, ヴィッツ, ヴォクシー, ヴェルファイア, ウィッシュ, フィット, フリード, ステップワゴン, セレナ, キューブ, ノート, モコ, デミオ, スイフト, ワゴン R, ムーヴ, ミラ, タント, eK

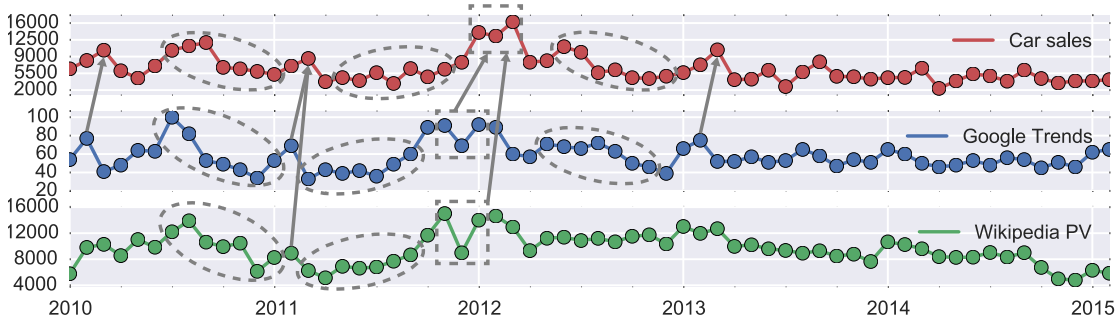


図 1: ホンダ・フリードの販売台数及び対応する Google Trends 値、Wikipedia 閲覧数の推移

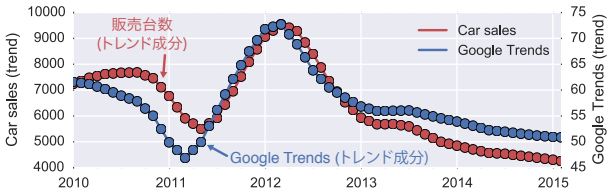


図 2: ホンダ・フリードの販売台数と Google Trends 値のトレンド成分

毎月の閲覧数を Wikimedia 閲覧数統計データ *6 から集計した *7。各ページの閲覧数はリダイレクトページ *8 の閲覧数も合算している。

3.2 検索行動量と販売台数間の相関

まず、データとして得られた 22 車種について、それぞれ対応する Google Trends 値及び Wikipedia 閲覧数と販売台数の間で月別の相関係数を計算した。相関係数を計算する期間は、5 章の実験における学習データ区間と同様に 2010 年 1 月〜2013 年 8 月とした。ここで、検索行動は購買行動に先立って行われることを考慮し、各検索行動量の系列を 1 ヶ月及び 2 ヶ月先行させた場合についても計算を行った。

相関係数を計算した結果、Google Trends の場合は 7 車種、Wikipedia 閲覧数の場合は 6 車種について、相関係数が 0.4 以上であった。更に、Google Trends を 1, 2 ヶ月先行させた場合はそれぞれ 10, 6 車種、Wikipedia 閲覧数の場合は 4, 2 車種について相関係数が 0.4 以上であった。これらの車種については販売台数と検索行動量との相関が高いことから、販売台数の予測に検索行動量を用いることで予測精度を向上させられる可能性がある。また、先行させる期間に注目すると、多くの車種について Google Trends に対しては 1 ヶ月先行させた場合、Wikipedia 閲覧数に対しては先行させない場合が最も相関が高いが、一部の車種については異なる期間の場合の方が高い相関を持つ。このため、予測モデルでは予め定めた期間だけ先行させるのではなく、複数の先行期間について最良の場合を判断できることが望ましい。

具体的な例として、ホンダ・フリードの販売台数と各検索行動量の推移を比較した図を図 1 に示す。短期的な視点で見ると、売り上げがピークを迎える毎年 3 月期の伸びについて、

*6 Page view statistics for Wikimedia projects: http://dumps.wikimedia.org/other/pagecounts-raw/

*7 但し、マツダ・デミオとダイハツ・ミラの二車種については、ある特定の日に前後の日の 100 倍以上の閲覧数が記録されているが、これらは異常値であると考えて前後 1 日の平均値を代わりに用いた。

*8 例えば「プリウス」へアクセスすると「トヨタ・プリウス」へと自動で転送されるため、「プリウス」の閲覧数も合算した。

表 1: 販売台数のトレンド成分と Google Trends 値のトレンド成分間において相関係数が最も高い 8 車種におけるトレンド成分間の相関係数（2010 年 1 月〜2013 年 8 月）

車種	Google Trends			Wikipedia PV		
	shift0	shift1	shift2	shift0	shift1	shift2
Demio	0.65	**0.67**	**0.67**	-0.12	**-0.10**	**-0.10**
eK	**0.84**	0.77	0.69	**0.62**	0.58	0.54
Freed	0.81	**0.85**	0.83	**0.50**	0.44	0.35
Mira	**0.64**	0.62	0.57	0.79	0.84	**0.89**
Move	**0.88**	0.84	0.80	**0.67**	0.66	0.64
Note	**0.97**	0.94	0.90	0.90	**0.92**	**0.92**
Passo	**0.90**	0.82	0.74	**0.27**	0.22	0.16
WagonR	**0.72**	0.71	0.69	0.19	**0.28**	0.36

Google Trends の多く、また Wikipedia 閲覧数の一部が 1〜2 ヶ月前に捉えていることが分かる。一方、長期的な視点で見ると、2011 年初頭までの下降・2011 年初頭から 2012 年初頭までの上昇・2012 年初頭からの下降トレンドをやや先行して捉えていることが分かる。

3.3 検索行動量と販売台数のトレンド成分間の相関

検索行動量からは毎年 3 月期のピークのような周期的な要因よりも、トレンドや単発的なピークを捉えられることが望ましい。この理由は、周期的要因は過去の販売台数系列からも自己回帰的に求めることが比較的容易なためである。

特にトレンド成分のみに焦点を合わせて先行性を詳しく見るため、販売台数と検索行動量を STL [Cleveland 90] により 12 期（12 ヶ月）の季節成分とトレンド成分に分解し、トレンド成分同士について観察を行った。具体的に、ホンダ・フリードの販売台数と Google Trends の各トレンド成分を比較した図を図 2 に示す。図を見ると、ホンダ・フリードの場合は 1〜2 ヶ月程度先行してトレンドの変化を捉えられていることが分かる。その他の車種も合わせた全 22 車種について、同様の分析を 0, 1, 2 ヶ月先行させた場合で行った結果、Google Trends の場合はそれぞれ 8, 8, 6 車種、Wikipedia 閲覧数の場合はそれぞれ 4, 5, 5 車種について、相関係数が 0.6 以上であった。このうち、最も Google Trends との相関係数が高かった 8 車種について、具体的な値を表 1 に示す。以上の観察より、いくつかの車種においては Google Trends や Wikipedia 閲覧数は、トレンドを捉える観点からも予測に有用な可能性を有すると言える。

4 提案手法

4.1 状態空間モデル

状態空間モデルは、ある内部状態からどのように観測値が生起するかを決定する観測方程式と、内部状態が時刻の経過

に従ってどのように変化するかを決定する状態方程式の 2 つから構成される。以下、本研究で用いるモデルについて、まず時刻 t における販売台数 y_t がどのような要因に分解されるか（観測方程式と対応する）を示し、続いて各要因がどのように遷移するか（状態方程式と対応する）を示す。

4.2 ベースラインモデル

まず、検索行動量を用いず、過去の時系列データのみから将来予測を行うベースラインモデルを示す。本研究では、自動車販売台数の要因として、短期的なトレンド要因と周期的な季節要因（12 ヶ月周期）の 2 つを仮定する。

具体的に、ある車種について時刻 t における自動車販売台数を y_t とした時、y_t を次のように分解する。

$$y_t = \mu_t + \gamma_{1,t} + v_t \tag{1}$$

ここで、μ_t は確率差分方程式 $\mu_t = 2\mu_{t-1} - \mu_{t-2} + w_{\mu_t}$ に従う 2 次のトレンド成分である。$\gamma_{1,t}$ は $\gamma_{1,t} = -\sum_{u=1}^{11} \gamma_{u,t-1} + w_{\gamma_t}$ に従う 12 ヶ月を周期とした季節成分である。なお、$\gamma_{i,t} (1 \leq i \leq 11)$ は、過去 11 ヶ月分の季節成分を保持する変数であり、$\gamma_{i,t} = \gamma_{i-1,t-1}$ である。$v_t, w_{\mu_t}, w_{\gamma_t}$ は誤差項であり、本研究では $v_t \sim N(0, V),\ w_{\mu_t} \sim N(0, W_\mu),\ w_{\gamma_t} \sim N(0, W_\gamma)$ とした。

4.3 提案モデル

次に、過去の時系列データに加え、検索行動量も併せて考慮して将来予測を行う提案モデルを示す。本モデルでは、販売台数の新たな要因として検索行動量系列のトレンドを仮定する。また、検索行動量系列には、ベースラインモデルと同じく 2 次のトレンド成分・12 ヶ月周期の季節成分を要因と仮定する。具体的には、自動車販売台数を y_t^{s1}、検索行動量を y_t^{s2} とした時、それぞれを次のように分解する。

$$y_t^{s1} = \mu_t^{s1} + \gamma_{1,t}^{s1} + \alpha\mu_t^{s2} + v_t^{s1} \tag{2}$$

$$y_t^{s2} = \mu_t^{s2} + \gamma_{1,t}^{s2} + v_t^{s2} \tag{3}$$

ここで、α は検索行動量トレンドの重みを決定するパラメータである。各 $\mu_t, \gamma_{1,t}, v_t$ の意味は式 (1) と同様であり、s_1, s_2 はそれぞれ対応する時系列を表す。

以上のモデルは、ある年月の自動車販売台数の予測において、同じ年月における検索行動量のトレンド成分を用いている。一方、3.3 節で観察したように、車種によっては検索行動量のトレンドが自動車販売台数のトレンドよりも先行して現れる場合もある。そこで、先行するトレンドを捉えるため、改変を加えた次の 2 つのバリエーションを考える。

バリエーションの 1 つは、式 (2) において、予め定めた m 期前のトレンドである μ_{t-m}^{s2} を記憶しておき、μ_t^{s2} の代わりに用いるモデルである。これにより、検索行動量のトレンドが m 期だけ先行していると言う仮定を考慮できる。

もう 1 つは、式 (2) において、異なる時刻におけるトレンドを同時に用いるモデルである。具体的に、0 期前から 2 期前のトレンドを同時に用いた場合、式 (2) は次のようになる。

$$y_t^{s1} = \mu_t^{s1} + \gamma_{1,t}^{s1} + \alpha_0\mu_t^{s2} + \alpha_1\mu_{-1,t}^{s2} + \alpha_2\mu_{-2,t}^{s2} + v_t^{s1} \tag{4}$$

ここで、$\mu_{-1,t}^{s2}, \mu_{-2,t}^{s2}$ はそれぞれ 1 期前、2 期前の検索行動量トレンド値に対応し、$\mu_{-1,t}^{s2} = \mu_{t-1}^{s2},\ \mu_{-2,t}^{s2} = \mu_{-1,t-1}^{s2}$ である。$\alpha_{0,1,2}$ はそれぞれ {0,1,2} 期前のトレンドの重みを決定する。

5 評価実験

5.1 実験条件

実験に用いた新車販売台数及び検索行動量のデータは 3.1 節で得たデータである。本研究では Google Trends 値を用いた改善が期待できる、表 1 の 8 車種に対する予測結果を示す。

実験で比較するベースラインは、4.2 節で説明した検索行動量を用いないモデルである（baseline）。また、提案手法は、一定の期間（0,1,2 期）シフトさせた検索行動量トレンドを用いるモデル（uni）に加え、1,2 期前までの（同時刻も含む）複数時刻におけるトレンドを同時に用いるモデル（multi）である。

モデルの各パラメータを決定する学習期間は 2010 年 1 月〜2013 年 8 月とし、予測精度を評価するテスト期間は 2013 年 9 月〜2015 年 2 月の 1.5 年間とした。各パラメータの計算には、R 言語のパッケージの一つである dlm 1.1-4 [*9] の最尤推定関数を用いた。誤差項 V, W_μ, W_γ の初期値について、分散の初期値には 10^7 を、共分散の初期値には 0 をそれぞれ用いた。

各月の予測値は、次のような手順で算出した。まず、モデルのパラメータを学習データに対する最尤推定により求め、モデル M を作る。次に、テスト期間中の各時刻 τ に対し、$\tau - n$ までのデータを用いた n 期先予測の予測値は次のように求める。

1. 訓練データの最初から $\tau - n$ までのデータを用い、モデル M を用いたカルマンフィルタで内部状態系列を求める
2. 求めた内部状態系列を利用し、n 期先である時刻 τ における新車販売台数を求める（状態方程式と観測方程式から求まる観測値分布の期待値）

評価指標には、実際の新車販売台数との RMS（Root Mean Square; 誤差の二乗和の平均）を用いた。但し、販売台数は車種によって大きく異なることから、誤差には真の値に対する予測値の比率に基づく相対的な値を用いた。なお、RMS は低いほど予測精度が高いことを意味する。

5.2 実験結果と考察

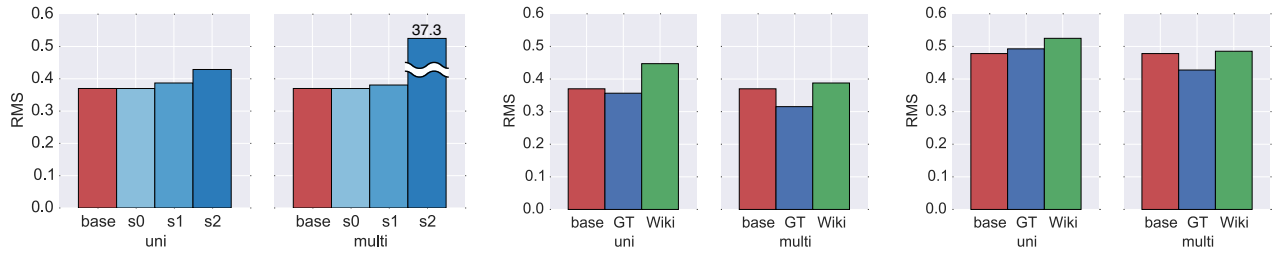
まず、図 3(a) に、ベースラインの RMS と提案モデルにおいてシフト期間を固定した場合の RMS を示す [*10]。なお、s0, s1, s2 は、それぞれシフト期間を示している。期間を固定した場合は、uni 及び multi モデルの双方で baseline モデルに対する性能差があまり出ておらず、期間をシフトさせた場合は若干予測精度が下がっている。これは、車種によって適切なシフト期間が異なるため、シフト期間を全ての車種で固定すると予測精度が改善・悪化する車種の双方が存在するためである。

具体的に、シフト期間を 0 ヶ月（シフトしない）に固定する uni-s0 の場合の予測精度を図 4 に示す。例えばホンダ・フリードや三菱・eK では RMS が軽減しているが、ダイハツ・ミラや日産・ノートでは悪化しているため、全体の平均で見た場合はベースラインとほぼ同じとなった。他のシフト期間の場合についても同様の傾向が見られた。

そこで、車種ごとに、販売数を予測する時刻より前までの区間において最も予測精度が高くなるシフト期間を選び、予測に利用した場合の結果を図 3(b)（1 期先予測）及び 3(c)（2 期先予測）に示す。検索行動量を用いない baseline（base）と Google Trends を用いる提案モデル（GT）を比較すると、1 期先予測の場合は uni 及び multi モデルの双方で多少の改善が見られ、特に multi モデルについては RMS が約 15% 改善

*9 Package dlm: `http://cran.r-project.org/web/packages/dlm/`

*10 図 3(a) において、multi-s2 モデルの予測精度が非常に悪くなっているのは、ある 1 車種について過適応してしまったためである。



(a) シフト期間固定・1 期先予測（Google Trend）　(b) シフト期間可変・1 期先予測　(c) シフト期間可変・2 期先予測

図 3: 各手法の予測精度（RMS）

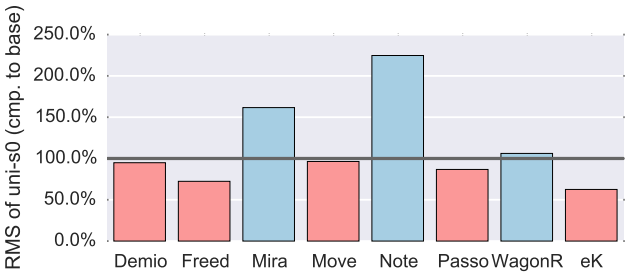


図 4: ベースラインと提案モデル uni-s0 の予測精度比較

した。2 期先予測の場合、uni モデルでは精度がやや悪化したが、multi モデルについては約 11% の改善が見られた。

一方、Google Trends の代わりに Wikipedia 閲覧数を用いた場合（Wiki）はベースラインよりもやや悪化する結果となった。この理由は、今回予測に用いた車種は Google Trends との相関が高い一方、Wikipedia 閲覧数との相関は必ずしも高くない車種であったため Wikipedia 閲覧数がノイズとして働いている車種があるためと考えている。

最後に、実際に三菱・eK の販売台数の uni-s1 モデルによる予測例を図 5 に示す。検索行動量を用いない baseline は、2014 年初頭の販売ピークを実際よりも低く見積もってしまっている。また、2014 年第 2 四半期に、販売ピークがあると予測しているが、実際にはそのようなピークは現れていない。これらは 2013 年の傾向と一致するため、過去の販売台数による影響が原因と考えている。これに対し、Google Trends を用いた uni-s1 では、Google Trends で観測された 2014 年初頭のピークに基いた予測が行えている。また、2014 年第 2 四半期については、Google Trends では大きな変化がなくトレンドも低調であることから、過度な見積もりを回避できている。

6 おわりに

本研究では、検索行動量と自動車販売台数について分析を行い、一部の車種において検索行動トレンドが販売台数より先行して現れることを確認した。更に、以上の分析に基いて検索行動のトレンドを考慮できる状態空間モデルを提案し、一部の車種について販売台数の将来予測精度が向上することを示した。

今後の課題として、以下の 2 点がある。まず、Wikipedia 閲覧数をより有効に使うための処理について検討したいと考えている。実験では Google Trends を用いた場合において予測精度を多少改善することができたが、Wikipedia 閲覧数を用いた場合は改善が見られなかった。検索行動量とみなすと言う観点で見た場合、Wikipedia 閲覧数は直接的な検索行動数を反映する Google Trends と比較すると別の要因によるノイズが加わるため、今後は Wikipedia 特有の要因について検討したい。また、ソーシャルメディアへの投稿についても考慮したいと考えている。例えば消費者の購買行動モデルの一つである AISAS® によると、購入前の検索（Search）に加え、購入後には情報共有（Share）が行われるとしており [現代用語の基礎知識 15]、口コミサイトなどへの投稿が将来予測に有用な可能性がある。検索行動と比較すると、消費行動よりも遅れて観察される点や、情報共有の内容によっては逆に消費行動が抑制される点などで異なるため、これらも考慮できるモデルを構築したい。

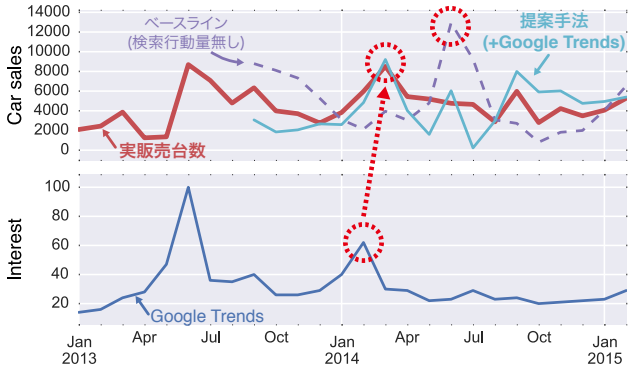


図 5: 提案モデル uni-s1 による三菱・eK の一期予測例と Google Trends 値の推移

参考文献

[Choi 12] Choi, H. and Varian, H.: Predicting the present with google trends, *Economic Record*, Vol. 88, No. s1, pp. 2–9 (2012)

[Cleveland 90] Cleveland, R. B., Cleveland, W. S., McRae, J. E., and Terpenning, I.: STL: A seasonal-trend decomposition procedure based on loess, *Journal of Official Statistics*, Vol. 6, No. 1, pp. 3–73 (1990)

[Goel 10] Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., and Watts, D. J.: Predicting consumer behavior with Web search, *Proceedings of the National Academy of Sciences*, Vol. 107, No. 41, pp. 17486–17490 (2010)

[Naik 99] Naik, P. A.: Estimating the Half-life of Advertisements, *Marketing Letters*, Vol. 10, No. 4, pp. 345–356 (1999)

[Xu 12] Xu, W., Li, Z., and Chen, Q.: Forecasting the unemployment rate by neural networks using search engine query data, in *45th Hawaii International Conference on System Science*, pp. 3591–3599 (2012)

[現代用語の基礎知識 15] 現代用語の基礎知識 JapanKnowledge Lib：AISAS（アイサス）(2015), http://japanknowledge.com/lib/display/?lid=5002013500830, 2015-03-10 参照

[本橋 12] 本橋 永至, 磯崎 直樹, 長尾 大道, 樋口 知之：状態空間モデルによるインターネット広告のクリック率予測, オペレーションズ・リサーチ：経営の科学, Vol. 57, No. 10, pp. 574–583 (2012)

[矢田 93] 矢田 健, 井上 正之, 北川 源四郎：カルマンフィルタによる通話料収入予測, 電子情報通信学会技術研究報告. IN, 情報ネットワーク, Vol. 93, No. 23, pp. 43–50 (1993)