慶應義塾大学学術情報リポジトリ
Keio Associated Repository of Academic resouces

Title	時系列解析手法を用いたお客様相談センターの着信件数の予測
Sub Title	
Author	蒲池, 直哉(Kamachi, Naoya) 林, 高樹(Hayashi, Takaki)
Publisher	慶應義塾大学大学院経営管理研究科
Publication year	2016
Jtitle	
JaLC DOI	
Abstract	
Notes	修士学位論文. 2016年度経営学 第3148号
Genre	Thesis or Dissertation
URL	https://koara.lib.keio.ac.jp/xoonips/modules/xoonips/detail.php?koara_id=KO40003001-00002016-3148

慶應義塾大学学術情報リポジトリ(KOARA)に掲載されているコンテンツの著作権は、それぞれの著作者、学会または出版社/発行者に帰属し、その権利は著作権法によって保護されています。引用にあたっては、著作権法を遵守してご利用ください。

The copyrights of content available on the KeiO Associated Repository of Academic resources (KOARA) belong to the respective authors, academic societies, or publishers/issuers, and these rights are protected by the Japanese Copyright Act. When quoting the content, please follow the Japanese copyright act.

慶應義塾大学大学院経営管理研究科修士課程

学位論文（　　2016　年度）

論文題名

時系列解析手法を用いたお客様相談センターの着信件数の予測

主　査	林 高樹　教授
副　査	高橋 大志　教授
副　査	市来嵜 治　専任講師
副　査	住田 潮　特任教授

氏　名	蒲池 直哉

目次

第 1 章

序論

本論文では、お客様相談センターへの問い合わせの着信件数の予測モデルの構築および考察を行う。以下の小節において、はじめにビジネス上でのデータ活用について述べ、それを踏まえ本研究の背景と意義及び目的について述べる。

1.1 はじめに

はじめに、ビジネスにおけるデータ活用の重要性とデータサイエンティストを取り巻く環境について述べる。IT のめまぐるしい発達によって、我々の身の回りは量・種類ともに無数のデータが存在しており、かつそれらが流動的に変化していくような時代を迎えている。このような環境の中、現在多くの企業は、これまでとりあえず蓄積していたようなデータを活用しようとしたり、また新たにデータを蓄積できるようなシステムを導入したりすることに躍起になっている。そのためデータサイエンティストという職業が登場し、大量のデータをビジネス活用していくことが求められるようになってきている。図 1.1 はビッグデータアナリティクスの市場規模の推移と予測を示している。2020 年は 2012 年の約 3 倍の市場規模になると予測されており、データサイエンティストへの期待が込められているのであろうと考えられる。データサイエンティストには、ビジネススキル、IT スキル、統計解析スキルの 3 つのスキル領域が必要とされており、すべてをスペシャリストレベルまで高めることが理想であるが難しいため、どれか一つの領域を自分のコアスキルとして持ちながら、他のスキルを磨いているという人が多い。そのような中、多くの企業がビジネスへのデータ活用の重要性に気づきデータサイエンティストの育成に取り組もとするものの、あまりうまくいっていないというのが現状である [1]。

文献 [1] によると、データサイエンティストの育成を阻む外的な要因として、ビジネスにおける分析スキルを蓄積する難しさがあると考えられている。その理由は 2 つあり、ビジネス

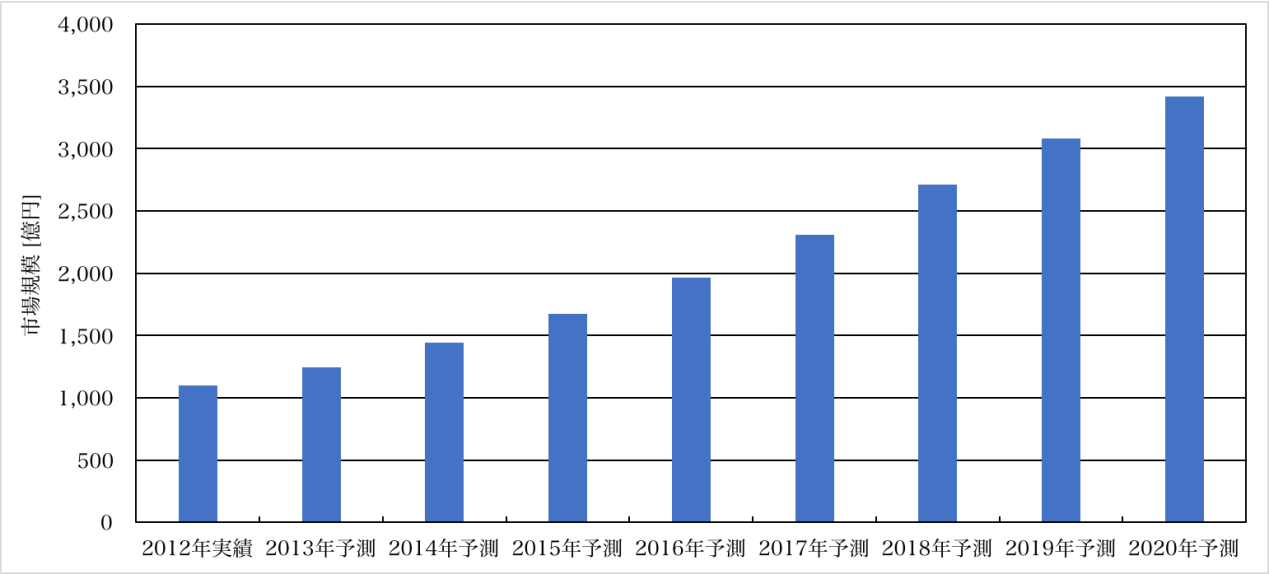

図 1.1　ビッグデータアナリティクスの市場規模推移と予測　　(出典：矢野経済研究所)

分析に関するまとまったノウハウがないこととビジネス分析を経験する機会が少ないことである。一つ目に関しては、実際のビジネスのデータに対して分析理論の仮定がうまく当てはまらなかったり、分析ツールの使い方がわからなかったり、そもそも何をやればいいのかわからないというような問題がある。また、データ分析にはデータのクリーニング（前処理）が必要であり、それが業務の大半を占める。そしていざ分析を行う際にも分析ツールが多いゆえに、それらの特徴を理解し上手く使い分けることが求められる。二つ目に関しては、データサイエンティストとして成長するには実務での経験が不可欠であるが、そもそも企業内で他の専門職と比べて割り当てられる人数が少なく、ビジネスの意思決定に関わることが多いため、新人が携わりにくいという環境であることも考えられる。以上のような外的な要因に加えて、現場で働いているデータサイエンティスト自体も問題を抱えていることもありうる。それは、データサイエンティストの「様々な分析手法を活用したい」という思いに対して、高度な分析手法から得られる結果がそのままビジネスで価値を発揮するとは限らないということである。ビジネスにおけるデータ分析の目的は、ビジネスで発生もしくは今後発生するであろう問題を発見し、統計解析や機械学習、データマイニングの各種方法論を駆使して、解決もしくは未然に防ぐことである。つまり、データ分析そのものの価値とは異なるということである。幾ら高度な分析を行ってもビジネスでの問題解決に繋がらなければ意味がないのである。しかし、シンプルな分析手法ばかり使用していてはデータサイエンティストとしてのスキル向上が期待できなくなることも考えられる。

このような現状において、データサイエンティストは、図 1.2 のようなフレームワークに

従ってビジネスでのデータ分析に取り組むことが有効であると考えられている [1]。このようなデータ分析フローに従うことで、解決すべき問題がずれたり、たとえ分析によって何か重要なことがわかっても打つべき施策がなかったりというようなビジネス価値の乏しいデータ分析となることを回避することができる。本研究においても、このビジネスにおけるデータ分析フローを意識しながら研究を進めていく。

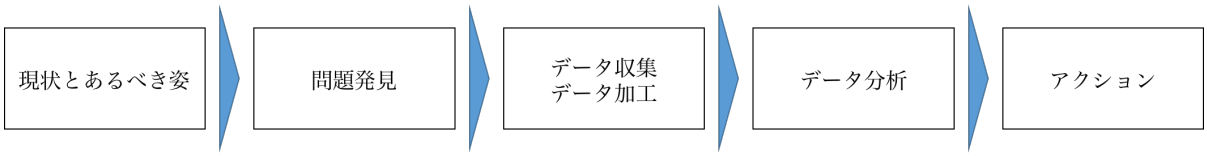

図 1.2　ビジネスにおけるデータ分析のフロー

1.2　背景と意義

本研究は、2016 年 4 月より本大学院である慶應義塾大学大学院経営管理研究科の学生と国内最大手の製造業である X 社の担当者とで共同研究の一環として行ってきたものである。X 社は主に住宅設備に関する商品を取り扱っており、我々の生活を支えている企業である。データの活用に積極的に取り組んでいる X 社との共同研究では、営業データ分析、アクセスログ解析、SNS データ解析などにも着手しているが、私はお客様相談センターの着信件数の予測について研究を行ってきた。

そこでまず、X 社が抱えているお客様相談センターについて述べる。X 社のお客様相談センターは、X 社で取り扱っている商品に関する相談や修理に関する問い合わせに対して対応を行っている。そのため、お客様相談センターのオペレーターには X 社の商品に関する知識をあらかじめ身につけることが求められる。また、X 社のお客様相談センターへの問い合わせは月に 10 万件以上も寄せられており、200 人を超えるオペレーターを外注で抱えている。そして、これらのオペレーターの手配については、X 社が提示する着信件数予測に基づいて体制を組んでいるという現状である。それ故に、着信件数予測の精度が良くないということは次のような 2 つの問題を引き起こす原因となりうる。

1. 実際よりも少なく予測をしてしまった場合は、受電率（対応できた件数 ÷ 実際かかってきた件数）が小さくなり、顧客満足度の低下に繋がる。
2. 実際よりも多く予測をしてしまった場合は、手待ちとなる人員が発生してしまい、人件費の増加に繋がる。

したがって、お客様相談センターの着信件数予測を行うことは X 社にとって意義あることであり、その予測精度を高めることで、上記のような問題を引き起こすことなく、外注しているオペレーターを効率良く育成し、かつ日々手配していくことができる体制を整えることに役立つことをを目指していきたいと考えている。

1.3　研究目的

本研究では、お客相談センターの着信件数の予測モデルを構築を行う。そこで、現在 X 社で実際に使用されてる予測モデルをベンチマークとして、さらに予測精度の高いモデルを構築することを本研究の目的とする。

1.4　本論文の構成

本論文は 6 つの章から構成される。第 2 章では理論として時系列解析の概念について述べる。続いて第 3 章では基礎分析を行い、お客様相談センターの着信件数の予測に関して X 社が取り組んできたことを先行事例としてとりあげ、現状の予測モデルの改善の余地について述べる。これらを踏まえ、第 4 章で予測モデルの構築を行い、第 5 章で従来のモデルとの予測精度の比較を行うことで、構築したモデルの考察をまとめる。最後に第 6 章で本研究の総括として結論を述べる。

第 2 章

時系列解析の方法論

本章では、文献 [2] 及び [3] に基づいて時系列解析に関する概念に関して述べる。

2.1　時系列とは

時系列解析について述べる上で、まずはじめに時系列に関して説明する。文献 [2] によると、『時系列 (time series) とは観測値 x_t の集合であり、各観測値はある特定の時間 t において記録されたものである』と定義されている。簡単に述べると、時間の経過とともに規則的あるいは不規則的に変動する現象を、連続的または不連続に記録したものが時系列である。我々の身近にある例を取り挙げると、気圧、気温や雨量などの気象情報、地震波の記録、GDP、株価や為替レートなどの経済現象の記録、血圧や脳波などの医学データなど、無数に存在する。これらの時系列データは観測される一定間隔によって、年次データ、月次データ、日時データなどと呼ばれる。また一般的に、時系列データの変動に関して以下の 4 つの要因が考えられている [4]。

1. 傾向変動（trend）　T_t
2. 循環変動（cycle）　C_t
3. 季節変動（seasonal）　S_t
4. 不規則変動（irregular）　I_t

そして、実際に観測される時系列データは上記の要素が重ね合わさってできたものである。具体的には、時系列 Y_t は以下のようなモデルとして表現することができると考えられている。

加法モデル：　$Y_t = T_t + C_t + S_t + I_t$

乗法モデル：　$Y_t = T_t \times C_t \times S_t \times I_t$

ただし、乗法モデルは時系列データおよび各要素が正の値をとると仮定することができる場合は、両辺に対数をとって以下のような加法モデルに書き換えることができる。

$$\log Y_t = \log T_t + \log C_t + \log S_t + \log I_t$$

時系列データがこのような要素の合成であれば、その各要素を抽出し、特定することによって解析に役立てることができる。

時系列解析における第一歩として重要なのは、まず時系列データを図示して見ることである。これによって、時系列の大まかな特徴を捉えることができるだけではなく、今後どのような解析を行うべきか、その方針を立てることも可能となる。そこで一例として、気象情報に関する時系列データについて考える。図 2.1 は直近 10 年間における東京の月の平均気温の時系列データである。グラフを見て分かるように、多少の誤差はあるものの一年間で 1 月から上昇し、8 月にピークを迎え、それ以降減少していくというような明確な年周期が現れており、これを 10 年間繰り返している。また解析としては、グラフからでは全体としての上昇あるいは下降トレンドは目視で確認できないのでトレンドの有無を調べたり、来年以降の予測を行うのであれば年周期をうまく取り入れたモデルを構築したり、などをあらかじめ考えることもできる。

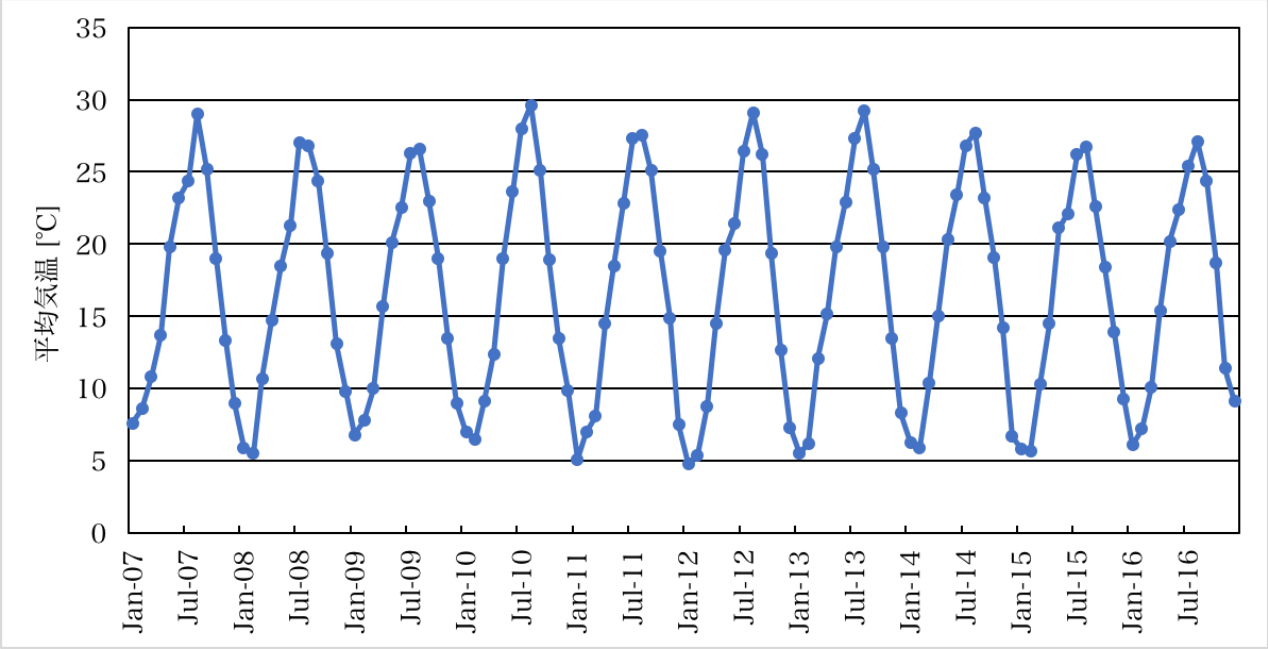

図 2.1　直近 10 年間における東京の月の平均気温の推移
(出典：『気象庁』の過去の気象データに基づき作成)

2.2　時系列データの分類

時系列データは様々な観点から分類することができる。以下にそれぞれの分類について述べる。

連続時間時系列と離散時間時系列

時間間隔の観点からはレコーダー等で連続的に記録されたデータは連続時間時系列と呼ばれるのに対し、1 時間おきに計測された気温のように、ある時間間隔で計測されたものは離散時間時系列と呼ばれる。さらに、離散時間時系列には等間隔に観測されたものと不等間隔なものがある。一般的に計算機等で解析を行う際には離散的な観測値を取り扱うことが多いので、離散時間時系列のことを単に時系列と呼ぶことが多い。

定常時系列と非定常時系列

時系列は時間とともに不規則な変動をしていることが多いが、時系列解析ではこのような不規則な変動を確率的なモデルを用いて表現することが重要となってくる。一見すると不規則な現象でも時間的に変化しない一定の確率的モデルの実現値とみなすことができる場合がある。このような時系列は定常 (stationary) 時系列と呼ばれる。時系列 Y_t が定常であるということは以下の 3 つの条件を満たすことを意味する。

1. すべての時点において、平均（期待値）が一定の同じ値である。

$$E(Y_t) = \mu \qquad \text{for } \forall \text{ t}$$

2. すべての時点において、分散が一定の同じ値である。

$$V(Y_t) = \sigma^2 \qquad \text{for } \forall \text{ t}$$

3. 2 つの時点の間の自己共分散 γ、自己相関係数 ρ が時間軸の絶対的な位置には依らず、2 時点の間隔 k が同じであれば同じ値となる。

$$Cov(Y_t, Y_{t-k}) = Cov(Y_s, Y_{s-k}) \equiv \gamma_k$$

$$\frac{Cov(Y_t, Y_{t-k})}{\sqrt{V(Y_t)}\sqrt{V(Y_{t-k})}} = \frac{Cov(Y_s, Y_{s-k})}{\sqrt{V(Y_s)}\sqrt{V(Y_{s-k})}} = \frac{\gamma_k}{\sigma^2} \equiv \rho_k$$

上記の条件は、正確には「弱定常性 (weakly stationary)」あるいは「2 次の定常性 (covariance stationary)」の条件である。ある時系列の分布が時間のシフトに関して不変で、その確率分布

を時間軸方向に移動しても変化しないとき、その時系列は強定常である。一方、平均のまわりの変動の仕方が時間的に変化しているものを非定常 (nonstationary) 時系列という。

ガウス型時系列と非ガウス型時系列

時系列の分布について、時系列の分布が正規分布に従うものはガウス型 (Gaussian) 時系列、そうでないものは非ガウス型 (non-Gaussian) 時系列である。モデルを構築する上で多くの場合はガウス分布に従うと仮定する。また、そのままでは正規分布に従うとはみなせない時系列でも、データに適切な変換を施すことによって、近似的にガウス型時系列とみなすこともある。

線形時系列と非線形時系列

モデルの観点からは、線形なモデルを出力として表現できるような時系列は線形 (linear) 時系列、非線形なモデルする必要がある時系列を非線形 (nonlinear) 時系列と呼ばれる。

欠損値と異常値

以上のような時系列の分類とは異なるが、時系列解析を行う上で注意するべきものとして欠損値 (missing value) と異常値 (outlier) が存在する。欠損値は何かしらの理由によって観測値が記録されなかった部分を示し、異常値は観測している現象自体の異常な振る舞い、観測装置の異常、記入やデータ転送時のミス等によって明らかに異常なデータの部分を示す。

2.3　時系列解析の流れ

前節のように、様々な観点から時系列を分類することができるが、時系列の統計的解析は大きく 4 つの観点に分けることができ、これらを総称して、時系列解析 (time series analysis) と考えられている。以下に、これらの 4 つの観点について簡単に説明する。

記述 (description)

記述とは時系列を図示したり、標本自己共分散関数、標本自己相関関数、ピリオドグラムなどの基本的な記述統計量を用いて時系列の特徴を簡潔に表現することである。時系列解析では多量の数値が出力されるので、グラフによって表現することが非常に効果的である。

信号抽出 (signal extraction)

信号抽出とは目的に応じて必要な信号や情報を取り出すことである。対象となる時系列の特徴や目的に応じて適切なモデリングを行うためには重要である。

モデリング (modeling)

モデリングとは与えられた時系列に対して、その変動の仕方を表現する時系列モデルを構築し、その時系列の確率モデルを解析することである。時系列には様々な特徴を持つものがあるので、解析の対象や目的に応じて適切な時系列モデルを選択し、そのモデルに含まれるパラメータを推定することが必要である。

予測 (prediction)

予測とは時系列が互いに相関を持つことを利用して、過去から現在までに与えらえた情報に基づいて将来の変動を予測することである。特に、モデリングによって推定されたモデルを利用して予測を行う場合が多い。

本研究ではモデルを構築し、最終的には将来の予測を行うことを目的としているため、上記の 4 つの観点を適宜考慮しながら、時系列解析を行っていく。

第 3 章

基礎分析と先行事例

2.1 節において、本研究で使用するデータの概要について述べる。これらのデータをもとに基礎分析の結果を 2.2 節で示し、2.3 節ではお客様相談センターの着信件数の予測に関して X 社が取り組んできたことを先行事例として取り挙げる。2.2 節、2.3 節の内容を踏まえ、2.4 節では現状の予測モデルの課題について検討を行う。

3.1　データ概要の説明

本研究で使用するデータについて述べる。データの期間は 2012 年 12 月 1 日から 2016 年 12 月 13 日までとなっており、着信件数は日次データとなっている。着信件数に関しては、9 つの商品カテゴリ（トイレ、住設、浴室、キッチン、サッシドア、ビル、インテリア、EXT(エクステリア)、タイル）毎に着信件数が記録されている。また、着信件数に対して実際に対応できた件数の割合を意味する受電率に関しても同様に 9 つの商品カテゴリ毎に日次データとして記録されている。ただし、受電率に関しては 2014 年 10 月 1 日より記録されている。その他に、祝日、3 連休後、お客様相談センターの休業日などの日付のカテゴリが各日付に対して設定されている。以下にデータ概要についてまとめる。

表 3.1　本研究で使用するデータの概要

データ期間	2012 年 12 月 1 日から 2016 年 12 月 13 日
データ形式	日次データ
着信件数のデータの商品カテゴリ	トイレ、住設、浴室、キッチン、サッシドア、ビル、インテリア、EXT、タイル
カテゴリ毎の受電率	対応できた件数 ÷ 着信件数
日付のカテゴリ	平日、土、日、祝日、3 連休後、休業日、休業日前、休業日後

3.2　基礎分析

まずはじめに商品カテゴリ毎の着信件数の違いについて確認する。図 3.1 は直近一年間における商品カテゴリ別の日時着信件数の平均の違いを示している。一日あたりの総着信件数はおよそ 4,000 件であり、そのうちトイレに関する着信件数が全体の約 4 割を占めている。この結果と X 社にとってトイレが主力商品であることも考慮して、以後本研究ではトイレに関する日次の着信件数について考えていく。

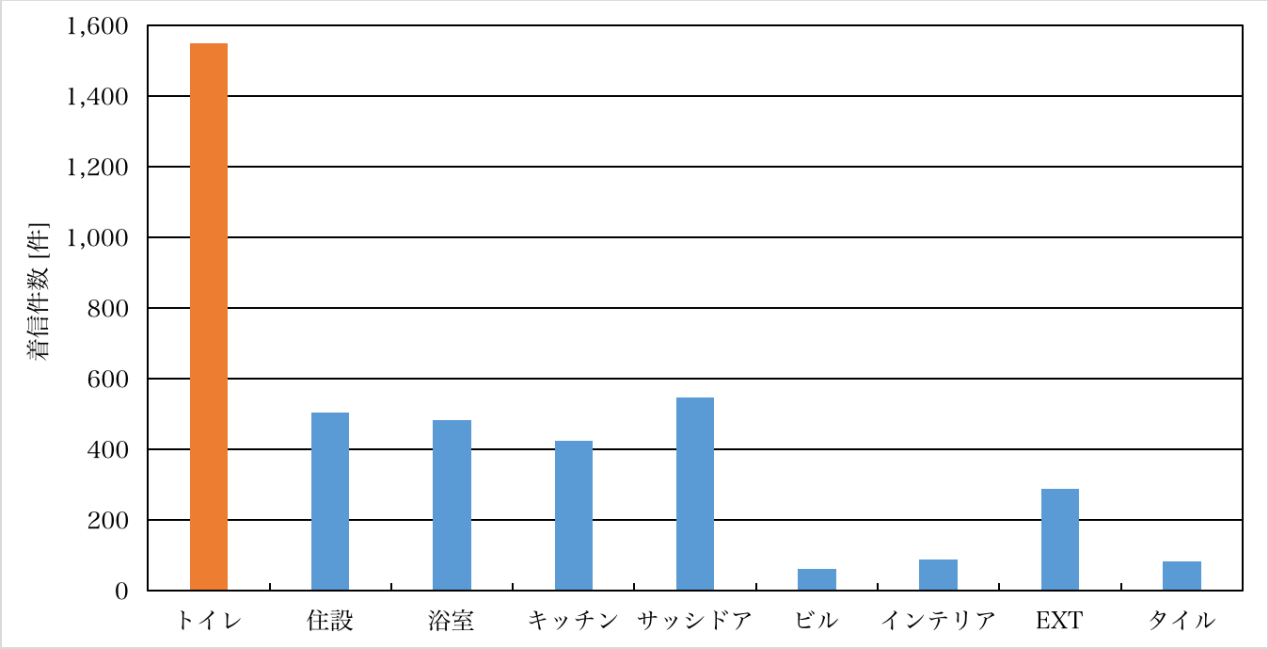

図 3.1　直近 1 年間における商品カテゴリ別平均日次着信件数

次に、時系列解析の第一歩として時系列の全体像の把握を行う。図 3.2 の推移から読み取れることとして、まず全体のトレンドとしては、2012 年から 2015 年 5 月あたりまでの期間では大小問わず目視で確認できるような上昇傾向も下降傾向も観測できないが、2015 年 6 月より、全体の推移の水準が上がっている。これは、2015 年 6 月より、本研究の研究対象である X 社のお客様相談センターにおいて、新たに水回りの修理に関する問い合わせも受け付けるようになったためである。また一年単位毎に見ると、毎年同じような変動を繰り返しており、その中で明確な季節変動も観測できる。大まかに述べると、3,4 月頃から上昇し、9 月頃から徐々に下降していくという季節変動である。異常値に関しては、お盆休みやゴールデンウィーク後に着信件数が大きく突出している箇所が観測される。一方、着信件数が 0 となっている部分に関してはお客様相談センターの休業日を意味している。また、このグラフからは読み取りづらい

かもしれないが、欠損値は存在していない。

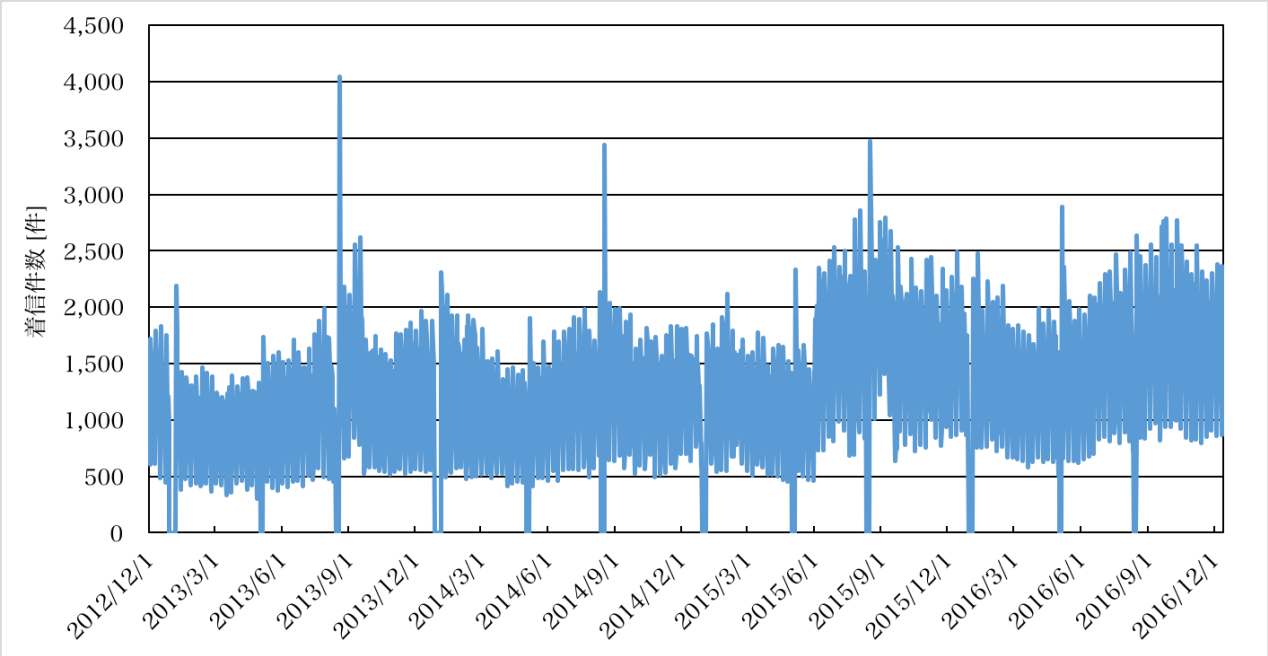

図 3.2　トイレに関する日次着信件数の推移

次は、図 3.2 の推移からは観測できないミクロな変動について確認する。図 3.3 は直近 6 ヶ月間におけるトイレに関する日次着信件数の推移を表しており、図 3.2 に比べてより細部まで時系列の変動を確認することができる。このグラフより、特徴的な波形が繰り返し現れていることがわかる。この特徴的な一つの波形は一週間の推移を表しており、それをよりわかりやすく図示したものが図 3.4 である。このグラフより、特徴的な波形は月曜日にピークを迎え、水曜日までは減少し、木曜日・金曜日は微増し、土曜日・日曜日は順々に減少するという一週間の変動によって形作られていることが読み取れる。

次は、図 3.5 において、設定されている日付のカテゴリによる着信件数の違いを確認する。図 3.5 は直近 1 年間における日付のカテゴリ別の日次着信件数の平均の違いを示している（ただし、休業日に関しては着信件数は 0 であるため除外している)。特徴としては、祝日は土日と同程度に少なく、休業日前は土日よりは多いが平日に比べると少ない水準となっている。一方、3 連休後や休業日後では平日に比べて高い水準となっていることが確認できる。

最後に、受電率に関する時系列の推移を確認する。図 3.6 は受電率の記録が開始された 2014 年 10 月から直近までのトイレに関する受電率の推移を示している。ただし、受電率が 0 と

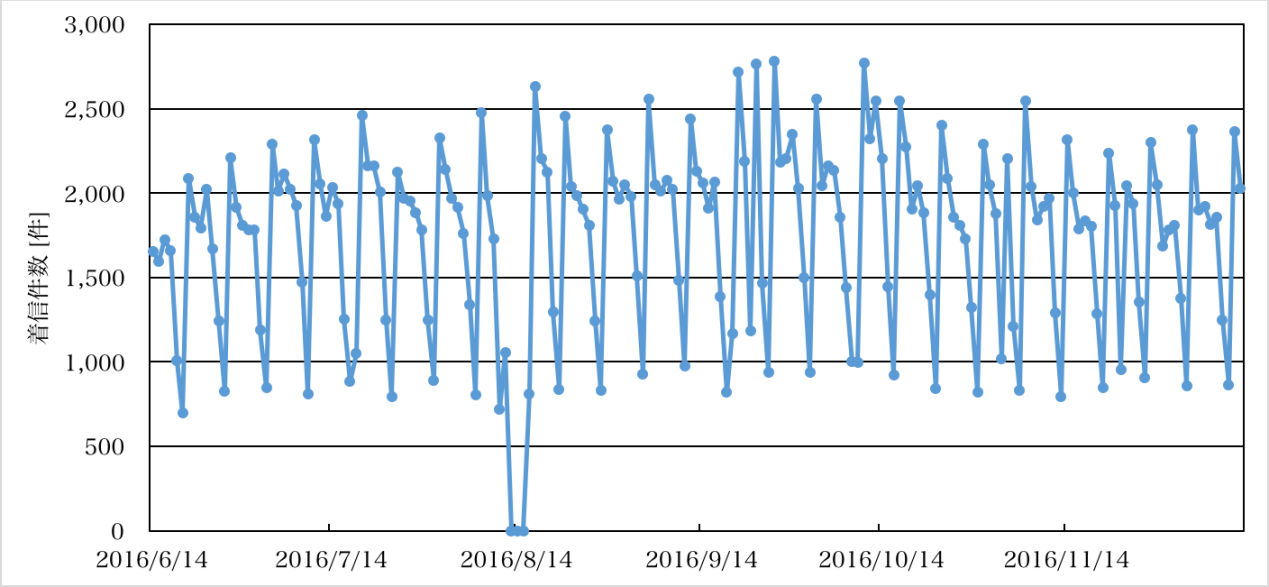

図 3.3　直近 6 ヶ月間におけるトイレに関する日次着信件数の推移

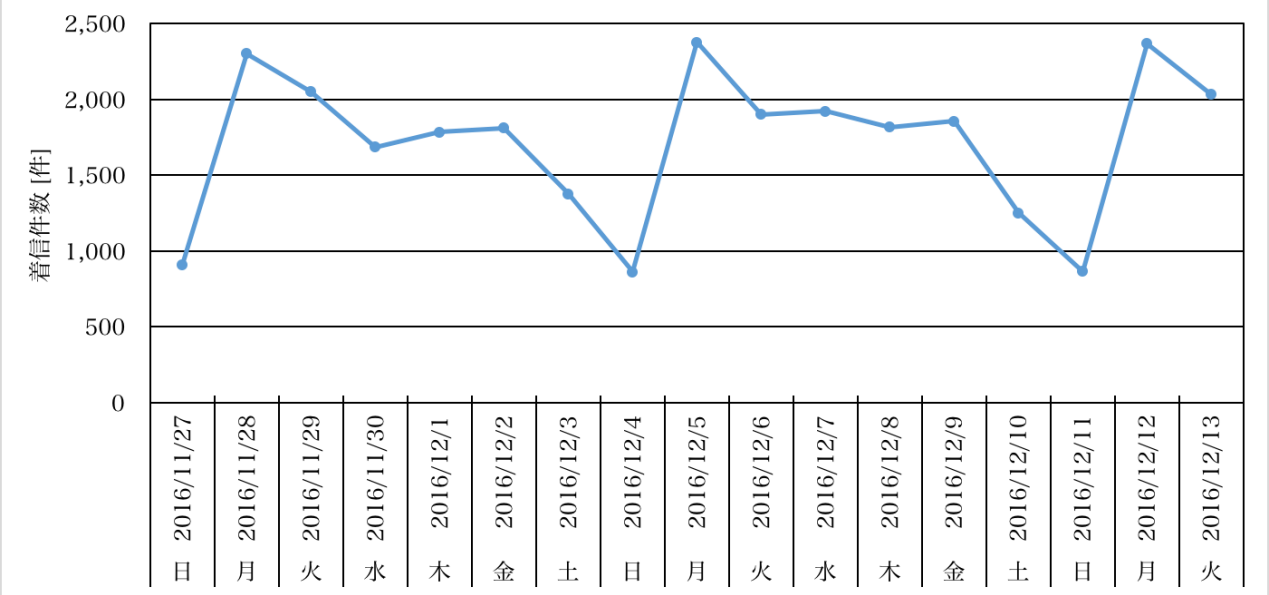

図 3.4　トイレに関する日次着信件数の曜日ごとの特徴

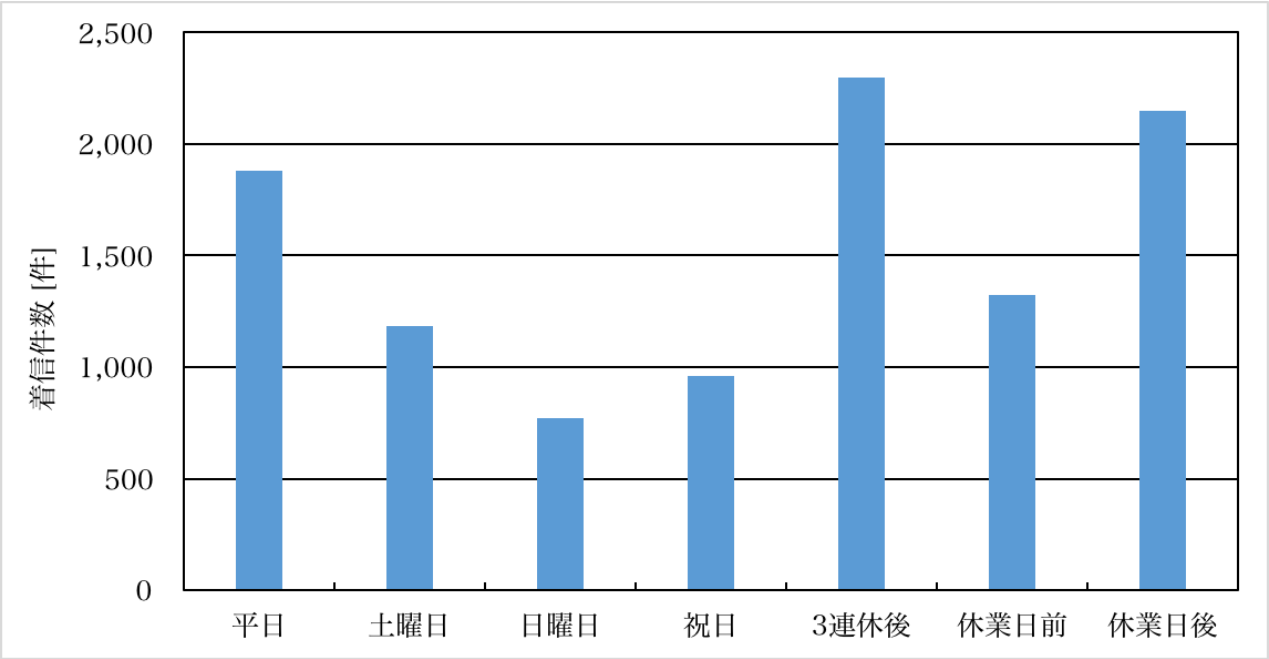

図 3.5　直近 1 年間における日付カテゴリ別の平均日次着信件数

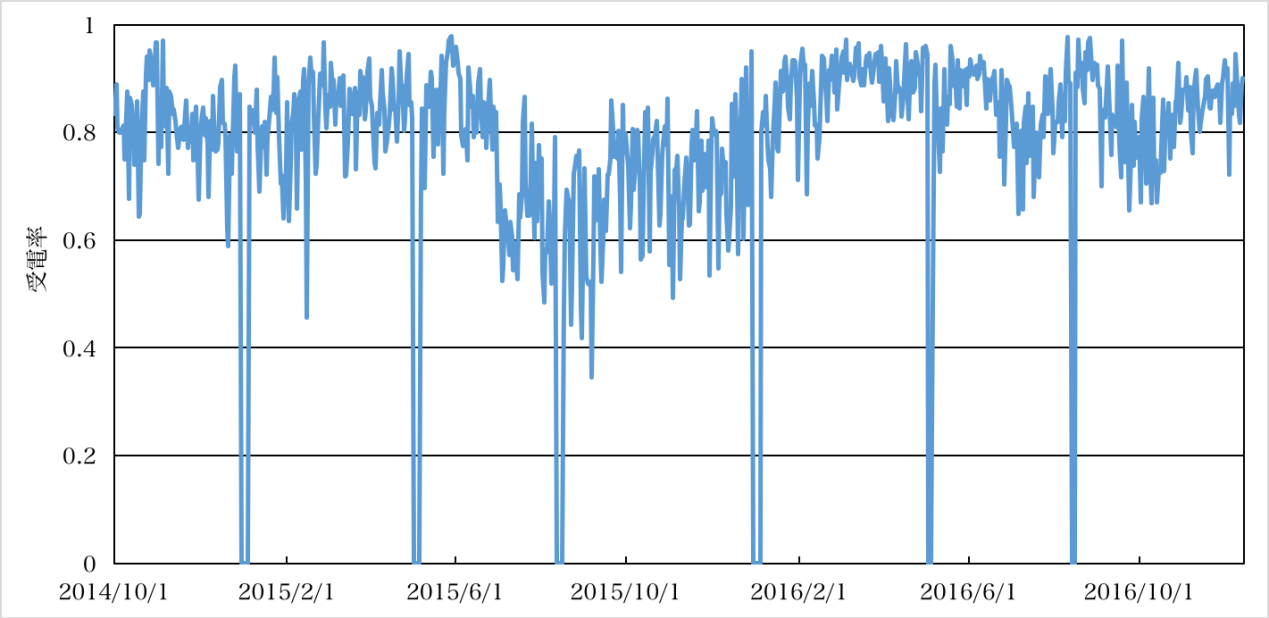

図 3.6　直近までのトイレに関する受電率の推移

なっている部分は着信件数と同様にお客様相談センターの休業日を意味している。図 3.6 を見るとわかるように、2015 年 6 月から 12 月の間は他の変動とは明らかに異なっている。2015 年に急に減少し、それ以降徐々に増加していることから、2015 年 6 月より開始された水回りの修理に関する問い合わせを受け付けるようになり、お客様相談センターにおける人員の配分体制に乱れが生じたためではないかと考えられる。

3.3　先行事例

お客様相談センターの着信件数の予測に関して、X 社の担当者が取り組んできたことを先行事例として述べる。まず、予測を行う上で複数のモデルを構築を行なっており、その中でも精度が良く、実際に採用されているモデルは大きく 2 つに分けることができる。1 つ目は ARIMA モデルであり、2 つ目は重回帰モデルである。以下、この 2 つのモデルに関する理論的な背景を説明を行なった上で、これらのモデルが実際にどのように使用されているかを述べる [3][4][5][6]。

3.3.1　ARIMA モデル

ARIMA モデルを説明するにあたって、はじめに AR モデル、MA モデルについて説明を行う。AR モデルの中で最も簡単な構造の線形定常確率モデルが「1 次の自己回帰モデル」AR(1) モデル (first-order autoregressive model) である。ある時系列を y_t を AR(1) モデルで表現すると、

$$y_t = m + \phi \times y_{t-1} + \epsilon_t, \qquad \epsilon_t \ \sim \ N(0, \sigma_\epsilon^2) \tag{3.1}$$

となる。ただし、m は定数項、ϕ は一期前の自身 y_{t-1} の係数、ϵ_t は平均 0, 標準偏差 σ_ϵ^2 の正規分布に従う誤差項である。この自己回帰モデルを一般化したものが AR(p) モデルであり、

$$y_t = m + \phi_1 \times y_{t-1} + \cdots + \phi_p \times y_{t-p} + \epsilon_t, \qquad \epsilon_t \ \sim \ N(0, \sigma_\epsilon^2) \tag{3.2}$$

と表現することができる。AR モデルには定常過程となるための条件がある。その条件は、特性方程式

$$\lambda^p - \phi_1 \lambda^{p-1} - \cdots - \phi_{p-1} \lambda = 0 \tag{3.3}$$

の解 $(\lambda_1, \lambda_2, \cdots, \lambda_p)$ の絶対値が 1 より小さいことである。一方、MA モデルについても同様に、最も簡単な構造のモデルが「1 次の移動平均モデル」MA(1) モデル (first-order moving average model) であり、時系列 y_t を当期 t と前期 $t-1$ の誤差項から成り立っており、

$$y_t = m + \epsilon_t + \theta \times \epsilon_{t-1}, \qquad \epsilon_t \ \sim \ N(0, \sigma_\epsilon^2) \tag{3.4}$$

と表現される。θ は時系列 y_t の前期 $t-1$ の誤差項 ϵ_{t-1} の係数である。AR モデルには上記のようなパラメータに定常過程となるための条件があるが、MA モデルはそのような条件なしに定常過程となる。そして、移動平均モデルを一般化したモデル MA(q) は

$$y_t = m + \epsilon_t + \theta_1 \times \epsilon_{t-1} + \cdots + \theta_q \times \epsilon_{t-q}, \qquad \epsilon_t \ \sim \ N(0, \sigma_\epsilon^2) \tag{3.5}$$

と表現できる。以上より、AR(p) モデルと MA(q) モデルを組み合わせたモデルを ARMA(p,q) モデル (自己回帰移動平均モデル：autoregressive moving average model) と言い、

$$y_t = m + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q}, \quad \epsilon_t \ \sim \ N(0, \sigma_\epsilon^2) \tag{3.6}$$

と表現できる。また、ラグ演算子を用いて表現することもできる。ラグ演算子とは、インプットの時系列に作用させることで、1 時点ずれた時系列をアウトプットとして返すことができるものである。ラグ演算子の記号を B(backward shift) とすると、

$$B y_t \equiv y_{t-1} \tag{3.7}$$

と定義できる。このラグ演算子 B を用いて ARMA(p, q) モデルを書き換えるために、

$$\begin{aligned} \phi(B) &\equiv 1 - \phi_1 B - \cdots - \phi_p B^p \\ \theta(B) &\equiv 1 + \theta_1 B + \cdots + \theta_q B^q \end{aligned} \tag{3.8}$$

と定義すると、(3.6) 式は次のように表現し直すことができる。

$$\phi(B) y_t = m + \theta(B) \epsilon_t \tag{3.9}$$

特性方程式 $z^p - \phi_1 z^{p-1} - \cdots - \phi_p = 0$ の解の絶対値が 1 より小さければ、$\phi(B)^{-1}$ が定義でき、(3.9) 式は、

$$\begin{aligned} y_t &= \phi(B)^{-1}(m + \theta(B)\epsilon_t) \\ &= \frac{m}{1 - \phi_1 - \cdots - \phi_p} + \frac{\theta(B)}{\phi(B)} \epsilon_t \end{aligned} \tag{3.10}$$

となる。また、ラグ演算子を用いると上記のように記述が簡潔になるだけでなく、様々な問題の解法に活用できる。例えば、時系列 y_t の平均は (3.10) 式より

$$E(y_t) = \frac{m}{1 - \phi_1 - \cdots - \phi_p} \equiv \mu \tag{3.11}$$

と求めることができる。ただし ARMA(p, q) モデルの場合、分散、自己共分散等はパラメータの簡単な式で表現できない (ARMA(1,1) モデルの場合は可能である)。

以上、AR、MA、ARMA モデルと順に説明してきたが、いずれも定常過程を表現する線形モデルである。しかし、実際のデータはトレンドを持つなど、そのままでは定常過程とみなすことができないことが多い。そこで、このような場合は時系列が定常過程とみなせるように変換することで ARMA モデル等が適用できるようにする。そうした変換のうち、最も主流で、簡単なものが「差分をとる」変換

$$\Delta y_t \equiv (1-B)y_t = y_t - y_{t-1}$$

である。仮に、時系列 y_t が直線的なトレンドを持つとし、以下のように定式化する。

$$y_t = (\alpha + \beta t) + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

この差分をとると、

$$\Delta y_t = \beta + (1-B)\sum_{j=0}^{\infty} \psi_j \epsilon_{t-j} = \beta + \psi_0 \epsilon_t + \sum_{j=1}^{\infty} (\psi_j - \psi_{j-1})\epsilon_{t-j}$$

とトレンド成分 (βt) が取り除かれる。また t^2 のトレンドを持つ場合は、1 回の差分では定常にならないが、次のように 2 回差分をとることで同様にトレンドを取り除くことができる。

$$y_t = (\alpha + \beta_1 t + \beta_2 t^2) + \sum_{j=0}^{\infty} \psi_j \epsilon_{t-j}$$

$$\Delta y_t = \beta_1 \beta_2 (2t-1) + \psi_0 \epsilon_t + \sum_{j=1}^{\infty} (\psi_j - \psi_{j-1})\epsilon_{t-j}$$

$$\Delta^2 y_t = 2\beta_2 + \psi_0 \epsilon_t + (\psi_1 - 2\psi_0)\epsilon_{t-1} + \sum_{j=2}^{\infty} (\psi_j - 2\psi_{j-1} + \psi_{j-2})\epsilon_{t-j}$$

したがって、一般的に d 回差分をとることで定常過程とすることができる確率過程を「d 次の和分過程」(d-th order integrated process) と呼び、d 回差分をとったものが ARMA(p,q) 過程を満たすものを ARIMA(p,d,q) 過程 (自己回帰和分移動平均過程：autoregressive integrated moving average process) という。この過程を満たす時系列 y_t を ARIMA(p,d,q) モデルで表現すると

$$\Delta^d y_t = m + \phi_1 \Delta^d y_{t-1} + \cdots + \phi_p \Delta^d y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots + \theta_q \epsilon_{t-q} \qquad (3.12)$$

となる。

また、季節変動のある時系列を ARIMA(p,d,q) の場合には、取り扱い方法は主に 3 つある。1 つ目は季節調整法である。この季節調整法によって季節変動を取り除いた時系列に対し

て ARIMA モデルを推定する。将来の予測を求める際は、季節変動の成分を補う必要があり、2.1 節で述べたように加法モデルならば季節要素を加え、乗法モデルならば季節要素を掛ける必要がある。2 つ目は季節階差をとる方法である。s 期前との差

$$\Delta_s y_t \equiv y_t - y_{t-s}$$

の系列を作り、これを ARIMA モデルに当てはめるという方法である。3 つ目は SARIMA モデルである。これは、季節変動についても ARIMA 構造を想定するモデルである。季節階差をとることはラグ演算子を用いて表現すると、

$$\Delta_s y_t = (1 - B^s) y_t$$

となる。l 回季節階差をとることを $\Delta_s^l y_t$ と表記する。この $\Delta_s^l y_t$ が 1 期前、2 期前などの同期の値の AR 過程に従うとすると、AR 過程を

$$\Phi_s(B) \equiv 1 - \phi_{s1} B^s - \phi_{s2} B^{2s} - \cdots - \phi_{sk} B^{ks}$$

として、

$$\Phi_s(B) \Delta_s^l y_t = \epsilon_t$$

となる。さらに、ϵ_t に関しても、1 期前、2 期前などの同期の値の MA 過程

$$\Theta_s(B) \equiv 1 + \theta_{s1} B^s + \theta_{s2} B^{2s} + \cdots + \theta_{sm} B^{ms}$$

に従うと仮定すれば、

$$\Phi_s(B) \Delta_s^l y_t = \Theta_s(B) \epsilon_t \tag{3.13}$$

というモデルを作ることができる。このモデルは、t、$t-s$、$t-2s$、・・・と 1 期ずつ隔てた変数の間の関係しか想定していないが、これにさらに通常の ARIMA(p,d,q) 過程を重ねることができる。

$$\Phi(B) \Phi_s(B) \Delta^d \Delta_s^l y_t = \Theta(B) \Theta_s(B) \epsilon_t \tag{3.14}$$

このように、通常の ARIMA(p,d,q) と季節階差に関する ARIMA(k,l,m) とを合わせたモデルを「季節 ARIMA(SARIMA：seasonal ARIMA) モデル」SARIMA$(p,d,q) \times (\mathrm{k,l,m})_S$ と呼ぶ。

実際、図 3.3、図 3.4 を見てわかるように、X 社のお客様相談センターの着信件数の時系列に関しても季節性がある。その季節性は週ごと、つまり曜日による明確な規則性が最も顕著である。そこで、ARIMA モデルを用いた現状の予測手法として主に 2 つある。1 つ目は上記で

も述べた季節階差をとる方法である。具体的には、7 日前との差 $(y_t - y_{t-7})$ をとることで週ごとの季節性を取り除き、それを ARIMA モデルに当てはめるという方法である。2 つ目は時系列データをあらかじめ曜日ごとにデータを分割することで、季節性を取り除く方法である。ARIMA モデルの特徴は、(3.12) 式からわかるように、t 時点での値を $t-1$、$t-2$、・・・のような自分自身の過去の値によって表現することである。さらに計算機を用いてモデルを構築する際に考える AR 項の次数 p はアルゴリズム上せいぜい 5 以下であるので、直近の値の線形結合によって決まる。したがって、予測を行う際には時系列の直近の傾向を上手く取り入れることができる反面、直近の現象が異常であれば、予測結果がその異常に影響を受ける可能性があると考えられる。

3.3.2　重回帰モデル

もう一つの予測モデルとして使用されているのが重回帰モデルである。重回帰モデルを説明する前に、説明変数が 1 つの場合である単回帰モデルについて述べる。単回帰モデルは、

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \qquad \epsilon_i \ \sim \ N(0, \sigma^2) \tag{3.15}$$

と表現できる。x_i を説明変数、y_i を被説明変数あるいは目的変数といい、あらかじめ観測、もしくは設定できる値 x_i に基づいて、目的となる変数 y_i の値を制御したり、予測したりすることが目的である。そこでまず、単回帰モデル (3.15) 式のパラメータ β_0 と β_1 を最小二乗法を用いた推定について述べる。パラメータ β_0 と β_1 の推定値を $\hat{\beta}_0$、$\hat{\beta}_1$、y_i の予測値を $\hat{y}_i$ とすると、

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{3.16}$$

と表される。そうすると、実測値 y_i と予測値 $\hat{y}_i$ の差である残差は、

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \tag{3.17}$$

となる。そして最小二乗法に基づいて残差平方和

$$S_e = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} \Big\{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \Big\}^2 \tag{3.18}$$

が最小になるような $\hat{\beta}_0$ と $\hat{\beta}_1$ を求める。そのために、S_e を $\hat{\beta}_0$ と $\hat{\beta}_1$ について偏微分したものを 0 とおき、整理すると以下のようになる。

$$\hat{\beta}_0 n + \hat{\beta}_1 \sum x_i = \sum y_i \tag{3.19}$$

$$\hat{\beta}_0 \sum x_i + \hat{\beta}_1 \sum x_i^2 = \sum x_i y_i \tag{3.20}$$

この (3.19) 式と (3.20) 式の連立方程式のことを正規方程式と呼ぶ。この正規方程式を解くと、

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1, \qquad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{3.21}$$

となる。ただし、S_{xy}、S_{xx} はそれぞれ x と y の偏差積和および x の平方和であり、

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}), \qquad S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.22}$$

と定義される。したがって、残差平方和の最小値は次のようになる。ただし、S_{yy} は y の平方和である。

$$S_e = S_{yy} - \hat{\beta}_1 S_{xy} \tag{3.23}$$

次に、推定された回帰モデルを評価するための寄与率および自由度調整済寄与率を説明する。まず、y の平方和 S_{yy} の分解を行う。

$$\begin{aligned} S_{yy} &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}\Big\{y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) + (\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\Big\}^2 \\ &\vdots \\ &= S_e + \sum_{i=1}^{n}\Big\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\Big\}^2 \end{aligned} \tag{3.24}$$

また、(3.23) 式より、$S_R = \hat{\beta}_1 S_{xy}$ とおくと、

$$S_{yy} = S_e + \hat{\beta}_1 S_{xy} = S_e + S_R \tag{3.25}$$

となる。上記の (3.24) 式と (3.25) 式を比較すると S_R は

$$S_R = \sum_{i=1}^{n}\Big\{(\hat{\beta}_0 + \hat{\beta}_1 x_i) - \bar{y}\Big\}^2 = \hat{\beta}_1 S_{xy} = \frac{S_{xy}^2}{S_{xx}} \tag{3.26}$$

と表せる。ここで S_{yy} は目的変数 y の平方和であり、y の全変動を意味し、一方 S_e は残差平方和であり、直線からのずれ具合を意味している。したがって (3.25) 式より、それらの平方和の差である S_R はデータの変動のうちで回帰直線によって説明できる部分を表すと考えることができる。実際に (3.26) 式の形より、S_R は推定された回帰式から決まる予測値 $\hat{y}_i$ が y の平均 $\bar{y}$ からどれくらいずれているかを測る量になっているので、S_R は回帰による平方和と呼ば

れる。この回帰による平方和 S_R の y の平方和に対する割合、つまり推定された回帰モデルの適合度を評価する指標である寄与率 R^2 は

$$R^2 = \frac{S_R}{S_{xy}} = 1 - \frac{S_e}{S_{yy}} \tag{3.27}$$

と定義できる。R^2 は全変動のうち回帰によって説明できる変動の割合を意味しており、1 に近いほど良い。また、寄与率 R^2 は x と y の相関係数 r_{xy} と次のような関係がある。

$$R^2 = \frac{S_R}{S_{yy}} = \frac{S_{xy}^2/S_{xx}}{S_{yy}} = \left(\frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}\right)^2 = r_{xy}^2 \tag{3.28}$$

さらに、(3.27) 式を自由度で調整すると、

$$R^{*2} = 1 - \frac{S_e/\phi_e}{S_{yy}/\phi_T} \tag{3.29}$$

となり、R^{*2} のことを自由度調整済寄与率と呼ぶ。ただし、ϕ_e は S_e の自由度、ϕ_T は S_{yy} の自由度である。一般的に重回帰モデルを考える場合に説明変数が多くなると寄与率 R^2 は 1 に近づく傾向にあるので、そのような場合は自由度調整済寄与率 R^{*2} によってモデルの説明力を判断するのが有効的となる。

次に重回帰モデルについて説明する。重回帰モデルは単回帰モデルを説明変数が 2 つ以上の場合に拡張したものであるが、単回帰モデルではなかった考え方や問題点が存在する。そこで説明変数が p 個の重回帰モデルを想定すると、

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \qquad \epsilon_i \ \sim \ N(0, \sigma^2) \tag{3.30}$$

のように表現することができる。単回帰モデルと同様に残差と残差平方和を

$$e_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}) \tag{3.31}$$

$$S_e = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left\{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}) \right\}^2 \tag{3.32}$$

と定義し、最小二乗法を用いて S_e を最小にするパラメータ $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ を求める。S_e を $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_p$ のそれぞれで偏微分したもの $\partial S_e/\partial \hat{\beta}_0 = 0, \partial S_e/\partial \hat{\beta}_1 = 0, \partial S_e/\partial \hat{\beta}_2 =$

$0, \cdots, \partial S_e/\partial \hat{\beta}_p = 0$ を整理すると以下のようになる。

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \hat{\beta}_2 \bar{x}_2 + \cdots + \hat{\beta}_p \bar{x}_p \tag{3.33}$$

$$\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1p} \\ S_{21} & S_{22} & \cdots & S_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ S_{p1} & S_{p2} & \cdots & S_{pp} \end{bmatrix}^{-1} \begin{bmatrix} S_{1y} \\ S_{2y} \\ \vdots \\ S_{py} \end{bmatrix} \tag{3.34}$$

ただし、平方和と偏差積和は次のように定義する。

$$S_{jk} = S_{kj} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \quad S_{jy} = \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(y_i - \bar{y}) \tag{3.35}$$

ここで、重回帰モデルで起こりうる問題点として (3.34) 式の逆行列がもとまらない場合である。このような場合を「多重共線性が存在する」と呼ぶ。多重共線性とは、説明変数 $x_1, x_2, \cdots, x_p$ のうちのいくつかの間に線形関係が成立していることである。このような場合には、線形関係を構成している（または構成していそうな）変数の一方を取り除いて解析をやり直す必要がある。また、サンプルサイズ n が説明変数の個数 p よりも小さい場合にも逆行列は存在しない。よって、残差平方和 S_e の最小値は次のようになる。

$$S_e = S_{xy} - (\hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + \cdots + \hat{\beta}_p S_{py}) \tag{3.36}$$

次に、重回帰モデルの寄与率と自由度調整済寄与率について説明する。単回帰モデルの場合と同様に y の平方和 S_{yy} の分解を行い、(3.36) 式と比較することで、回帰による平方和 S_R は

$$S_R = \hat{\beta}_1 S_{1y} + \hat{\beta}_2 S_{2y} + \cdots + \hat{\beta}_p S_{py} \tag{3.37}$$

と表せる。よって、寄与率と自由度調整済寄与率は以下のようになる。

$$R^2 = \frac{S_R}{S_{yy}} = 1 - \frac{S_e}{S_{yy}} \tag{3.38}$$

$$R^{*2} = 1 - \frac{S_e/\phi_e}{S_{yy}/\phi_T} \tag{3.39}$$

ただし、ϕ_e は S_e の自由度、ϕ_T は S_{yy} の自由度である。

また、単回帰モデルにはない重回帰モデルでは必要となる考え方として、説明変数の選択がある。変数の選択では、変数を入れ換えながら回帰モデルを構築し、より当てはまりの良いモデルを採択する。変数を選択することはモデルを選択することに等しく、モデルの選択とは複

数のモデルの中から真のモデルに近いモデルを見つけ出すことである。モデルの選択基準には、上記の自由度調整済寄与率 R^{*2}、回帰係数の $\Pr(>|t|)$ 値、F 値、AIC(赤池情報量基準) などがある。ただし、AIC は以下のように定義される。

$$AIC = -2 \times (\text{モデルの最大対数尤度}) + 2 \times (\text{モデルのパラメータ数}) \tag{3.40}$$

変数の選択の方法には、変数増減法、変数増加法、変数減少法などがある。

そこで、実際に予測に使用されている現状の重回帰モデルについて説明する。説明変数の候補は以下の表 3.2 の通りである。これらの説明変数はダミー変数 (0 or 1) として設定されてお

表 3.2　現状の重回帰モデルの説明変数の候補

説明変数の候補	内容
年	2012 年、2013 年、2014 年、2015 年、2016 年
月	1 月、2 月、3 月、 $\cdots$ 、11 月、12 月
曜日	月、火、水、 $\cdots$ 、土、日
受電率	対応できた件数 ÷ 着信件数
日付のカテゴリ	通常日 (平日, 土, 日)、祝日、3 連休後、休業日、休業日前、休業日後

り、これらの説明変数から変数増減法を用いてモデルの構築を行なっている。ARIMA モデルが過去の着信件数のみの線形結合で表現されるのに対し、重回帰モデルは着信件数を様々な要因 (説明変数) を用いて表現しているため、将来の予測を行うには想定できる要因による変動にはある程度対応できると考えられる。その反面、想定外の要因や環境が急な変化への対応が困難であると言える。

3.4　現状のモデルにおける課題

基礎分析および先行事例から現状の予測モデルの課題を抽出する。その課題を以下に 4 つ設定し、それらを順に説明する。

課題 1：　ARIMA モデルでは、祝日や 3 連休後などの特殊な日の予測に対応できていない。

課題 2：　ARIMA モデルでは、環境の変化により定常性が失われている。

課題 3：　重回帰モデルでは、誤差項が制約条件を満たしていない。

課題 4：　重回帰モデルの説明変数の追加および見直しの必要がある。

まず課題 1 について、現状の ARIMA モデルにおいては週毎による季節性、つまり着信件数の曜日による特徴のみを考慮している。また、予測を行う上で学習データ (予測以前のデータ) の特殊な日の数値に関しては、そのデータが該当する年月の曜日の平均値に修正した上で ARIMA モデルに当てはめている。具体的に述べると、例えば 2016 年 3 月 21 日 (月) の日付のカテゴリは祝日となっているので、その日の着信件数は 2016 年 3 月の祝日等でないの月曜日の平均値に修正を行なっている (以後、上記のデータの修正作業のことを「データ修正 A」と呼ぶ)。そのため、曜日による特徴は綺麗に捉えることができるが、実際には基礎分析の結果の図 3.5 からもわかるように日付のカテゴリ毎に着信件数が大きく異なっているので、祝日や 3 連休後のような特殊な日の予測には対応できていないという問題点がある。現在、X 社では ARIMA モデルの結果に対して、モデル外で特殊な日の効果を経験的に補っている。したがって、曜日のみならず日付のカテゴリ等も考慮したモデルを構築する必要がある。

次に課題 2 について、現状の 2 つの ARIMA モデルでは、階差をとったり、データを分割したりすることで季節性及びトレンドは取り除くことができるが、それでも定常性が完全ではない。例えば、図 3.7 は 7 日前との階差をとることで季節性を取り除いた時系列の推移を表している。この場合は、トレンドもなく、全体的には定常的な時系列となっているように見えるが、2015 年 6 月時点の急激な変化が残留している。一方、図 3.8 は月曜日のトイレに関する日次着信件数の推移を表しているが（ただし祝日等の特殊な日の着信件数はデータ修正 A を施している)、2015 年 6 月以降よりデータ全体の水準が上がっており、その時系列は定常性が保たれていないことがわかる。繰り返しになるが、これは 2015 年 6 月から新たに水回りの修理に関する問い合わせもこのお客様相談センターで受け付けるようになったためである。この場合は、再び階差をとる（例えば一期前との階差をとる）ことで、2015 年 6 月以降のデータ全体の水準の変化を取り除くことができるが、そのモデル構築までの工数の多さの割には、2015 年 6 月時点の急激な変化は残留したままである。したがって、曜日毎に時系列データを分割した ARIMA モデルは外部環境の変化によって定常性を失われてしまうため、このような変化にも対応でき、モデルとしても煩雑でないようなモデルを構築することが求められる。

次に課題 3 について、重回帰モデルを考える際に、そのモデル (3.30) の誤差項 ϵ_t には一般的に次のような仮定が存在する。

仮定 1：　すべての時点において、誤差項の期待値は 0。

$$E(\epsilon_t) = 0 \quad \text{for } \forall \text{ t}$$

仮定 2：　すべての時点において、誤差項の分散は等しい。(分散均一性：homoskedasticity)

$$V(\epsilon_t) = \sigma^2 \quad \text{for } \forall \text{ t}$$

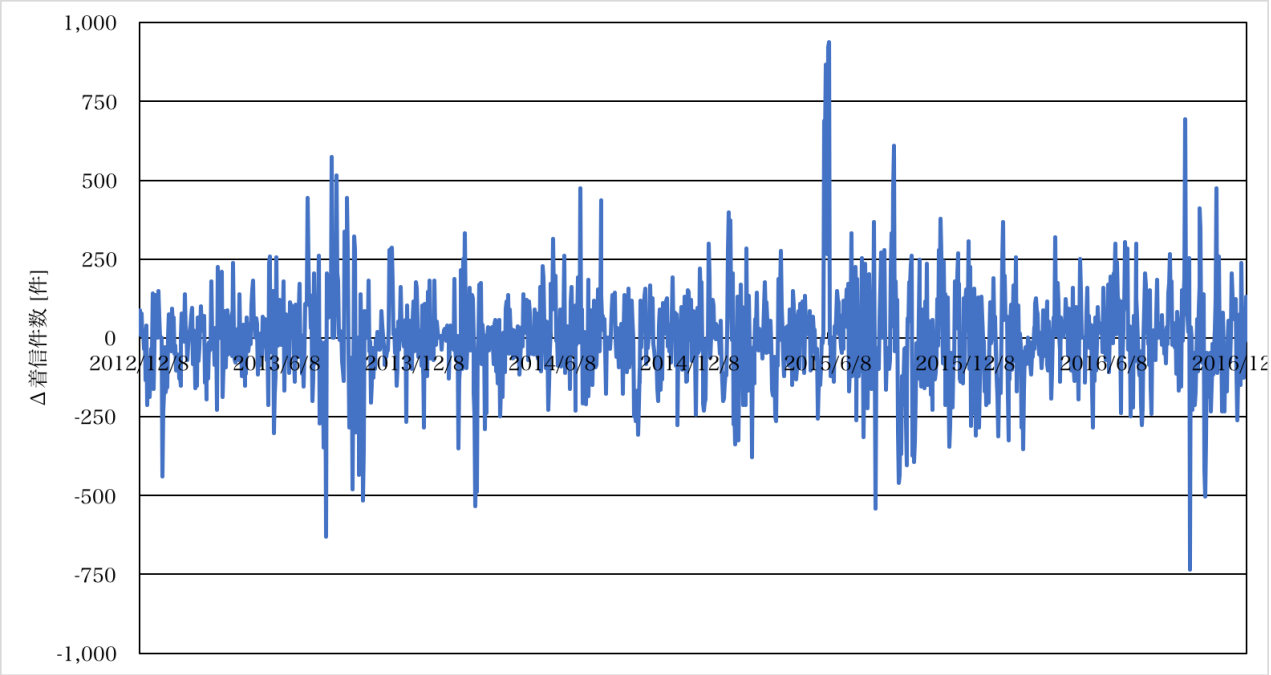

図 3.7　階差をとることで季節性を取り除いた時系列の推移

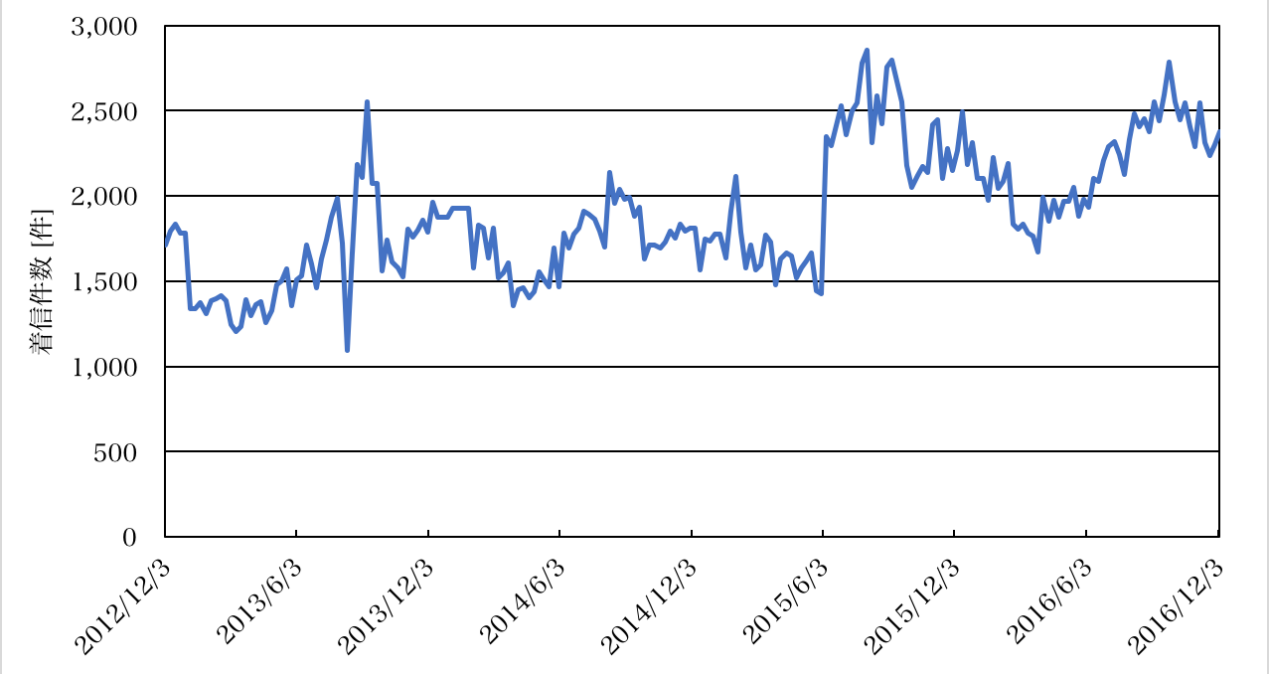

図 3.8　月曜日のトイレに関する日次着信件数の時系列の推移

仮定 3：　誤差項 ϵ_i と ϵ_j の共分散は 0。（誤差項の系列相関なし）

$$Cov(\epsilon_i, \epsilon_j) = 0 \quad \text{for } \forall \text{ i,j} \in \text{i} \neq \text{j}$$

仮定 4：　仮定 1〜3 と合わせると誤差項は互いに独立で，同一の正規分布に従う。

$$\epsilon_t \ \sim \ N(0, \sigma^2) \quad i.i.d.$$

これらの 4 つの仮定のうち、仮定 2、3 について着目する。そこで、2016 年 12 月 13 日までのデータ（ただし、休業日のデータは除く）に関して、説明変数を表 3.2 に則って設定した重回帰モデルを適用し、その誤差項の分散不均一性と系列相関の有無について検証する [7]。まず、誤差項の分散不均一性の有無を確認するために、Breusch-Pagan Test(ブルーシュ・ペイガン検定) を行う。Breusch-Pagan Test とは、誤差項の 2 乗値をもとの説明変数を用いて表した回帰式

$$\hat{\epsilon}_i^2 = \gamma_0 + \gamma_1 x_{i1} + \gamma_2 x_{i2} + \cdots + \gamma_p + v_i \tag{3.41}$$

を構築したときの当てはまり具合 (寄与率 R^2) を利用して、分散不均一性の有無を検証する方法である。帰無仮説 H_0 は誤差項は均一分散である、つまり、

$$H_0 \ : \ \gamma_1 = \gamma_2 = \cdots = \gamma_p = 0 \tag{3.42}$$

を満たすことである。対立仮説 H_1 は H_0 でない、つまり誤差項は不均一分散である。Breusch-Pagan Test で用いる検定統計量はサンプルサイズ n と (3.41) 式の寄与率 R^2 の積である nR^2 であり、これは自由度 p((3.41) 式の説明変数の数) の χ^2 分布に従う。そこで、有意水準 5% で χ^2 分布の上側臨界値を求め、検定統計量がこの臨界値より大きい値のとき H_0 を棄却する。そこで、実際に Breusch-Pagan Test を行なった結果が次の通りである。

```
        studentized Breusch-Pagan test

data:  reg.test
BP = 175.11, df = 25, p-value < 2.2e-16
```

図 3.9　Breusch-Pagan Test の結果

したがって、図 3.9 の結果より、p 値が 0.05 より小さく、帰無仮説 H_0 は棄却されるため、誤差項には不均一分散が存在しているということがわかる。次に、誤差項の系列相関の有無を確認するために、Durbin-Watson test(ダービン・ワトソン検定) を行う。Durbin-Watson test とは、誤差項が AR(1) 過程に従うと仮定し、

$$\epsilon_t = \rho \epsilon_{t-1} + v_t \tag{3.43}$$

と表した時に ρ が 0 か否かを確認することで、系列相関の有無を検証する方法である。帰無仮説 H_0 と対立仮説 H_1 はそれぞれ以下のようになる。

$$\begin{aligned} H_0 &: \rho = 0 \ (\text{系列相関が存在しない}) \\ H_1 &: \rho \neq 0 \ (\text{系列相関が存在する}) \end{aligned} \tag{3.44}$$

Durbin-Watson test で用いる検定統計量はダービン・ワトソン DW 比であり、以下のように定義される。

$$DW = \frac{\sum_{t=2}^{T}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{T}\epsilon_t^2} \tag{3.45}$$

この DW 比は説明変数の数やサンプルサイズによる違いもあるが、一般的に 2 に近いほど良いとされている。一方、DW 比が 2 からかけ離れ、0 に近いほど正の相関があり、4 に近いほど負の相関があるとされている。そこで、実際に Durbin-Watson test を行なった結果が次の通りである。

```
        Durbin-Watson test

data:  reg.test3
DW = 1.1277, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

図 3.10　Durbin-Watson test の結果

したがって図 3.10 の結果より、DW も 2 からかけ離れ 1 に近く、かつ p 値が 0.05 より小さいので、帰無仮説 H_0 は棄却され、誤差項には系列相関 (正の系列相関) があることがわかる。以上のことから、現状の重回帰モデルは誤差項には分散不均一性と系列相関が存在し仮定に反しているため、このような性質を考慮したモデルを構築するべきである。

最後に課題 4 について、現状のモデルで使用されている説明変数の候補は表 3.2 に示した通りである。これらの他に、時系列データの特徴を表している指標が存在しないかを検討する必要性がある。また、モデルを構築する前段階として、設定されている説明変数、特に日付のカテゴリの分類が妥当であるかを確認するべきである。もし仮に設定していた日付カテゴリと異なる挙動を示しているデータがある場合は、そのデータを予測のための学習データの一部として使用する際には、日付カテゴリを変更しておくべきであると考えられる。これによって、モデルの適正さが向上することが期待できる。

第 4 章

予測モデルの構築

前章で設定した課題の解決に向けて、予測モデルの構築を行う。その前段階として、課題 4 を解決するためにデータ全体の加工処理および説明変数の検討を行う。それに基づいて、本研究では 3 つのモデル案を提唱する。

4.1　予測モデル構築の準備

予測モデルの構築を行う前に、データの加工処理および説明変数の検討を行う。まず、データの加工処理を行う。具体的には、設定されている日付カテゴリの変更の検討である。そこで、各日付のカテゴリの説明を順に行う。通常日とは、祝日や 3 連休後などではない月曜日、火曜日、・・・、日曜日を意味している。次に祝日とは、いわゆる日本の暦上で設定されている祝日のうち、X 社のお客様相談センターが営業している日を意味している。そして 3 連休後とは、「土曜日→日曜日→祝日」もしくは「祝日→土曜日→日曜日」の翌日を意味している。休業日は年末年始、ゴールデンウィークやお盆休みの連続的な期間であり、お客様相談センターが営業していない日を意味している。よって、以後休業日は予測モデルの構築の対象データから取り除いて考える。そして、休業日前は休業日の前日を意味している（休業日の前々日も休業日前と設定されている場合もある）。最後に休業日後とは、休業日の期間が終わり、翌日からその週の金曜日までの期間を連休後と設定している。具体的な例を挙げると、2016 年のお盆休みは 8 月 11 日 (木) から 16 日 (火) までであり、翌日 17 日 (水) から 19 日 (金) までが「休業日後」と設定されている。これを踏まえ、日付のカテゴリの通常日に対する変化率の平均値と標準偏差をまとめると表 4.1 のようになる。ただし、変化率を測定する際は、該当する日付カテゴリが「通常日」である曜日の直近 1 ヶ月の平均に対する変化率を考える。この表 4.1 より、着信件数は祝日では通常日の半分程度、休業日前では約 7 割に減少している。一方、3 連休後、休業日後では通常日の約 2 割増しとなっている。これらの結果に基づいて、以下のように日付のカテゴリの変更を実施する。ただしデータ内には、本来ならば「休業日前」である

表 4.1　各日付カテゴリの通常日に対する変化率の平均値と標準偏差の比較

	祝日	3 連休後	休業日前	休業日後
平均値	40.6%	118.6%	71.0%	121.8%
標準偏差	10.1%	12.6%	16.5%	23.6%

表 4.2　日付のカテゴリの変更

日付	変更内容
2013 年 8 月 12 日	「通常日 (平日)」から「休業日前」に変更
2015 年 8 月 10 日	「通常日 (平日)」から「休業日前」に変更
2015 年 9 月 19 日	「通常日 (土)」から「連休中 (祝日と同等扱い)」に変更
2015 年 9 月 20 日	「通常日 (日)」から「連休中 (祝日と同等扱い)」に変更
2015 年 9 月 24 日	「休業日後」から「連休後 (3 連休後と同等扱い)」に変更
2015 年 9 月 25 日	「休業日後」から「通常日 (平日)」に変更
2016 年 4 月 30 日	「通常日 (土)」から「休業日前」に変更
2016 年 5 月 1 日	「通常日 (日)」から「休業日前」に変更
2016 年 5 月 2 日	「3 連休後」から「3 連休後かつ休業日前」に変更
2016 年 8 月 10 日	「通常日 (平日)」から「休業日前」に変更
2016 年 9 月 23 日	「通常日 (平日)」から「連休後 (3 連休後と同等扱い)」に変更

が「通常日 (平日)」へと既に変更されている箇所も存在する。これに関しては、表 4.1 の結果から判断して、既に日付のカテゴリが変更されていると考えられるのでこのままの状態にしておく。また、表 4.2 を見てわかるように、新たに「連休中」、「連休後」という日付カテゴリを設定したが、変化率の結果を考慮して連休中は祝日と、連休後は 3 連休後と同等の扱いをする。以後、このようなデータの日付のカテゴリの変更を行う作業のことを「データ修正 B」と呼ぶ。次節以降、予測モデルを構築する上で学習データに関してデータ修正 B を施した上でモデルに当てはめる。

次に、説明変数の検討を行う。まず、説明変数の候補として追加するべきであると考えられるのは、図 3.2 をはじめ、様々な場面で予測モデル構築の障害になっている、2015 年 6 月より開始された水回りの修理に関する問い合わせ受け付けるようになったことである。よって、「新規の修理に関する問い合わせの有無」を説明変数の候補として追加する。また表 4.1 の結果を見ると、休業日後に関しては変化率の標準偏差が他の日付のカテゴリと比べて高いことが読み取れる。このことから考えられるのは、同じ「休業日後」という日付のカテゴリでも、休

業日後の 1 日目と 4 日目とでは、変化率が異なるのではないかということである。そこで、休業日後の経過日数ごとに変化率の平均値と標準偏差を求めると以下の通りである。

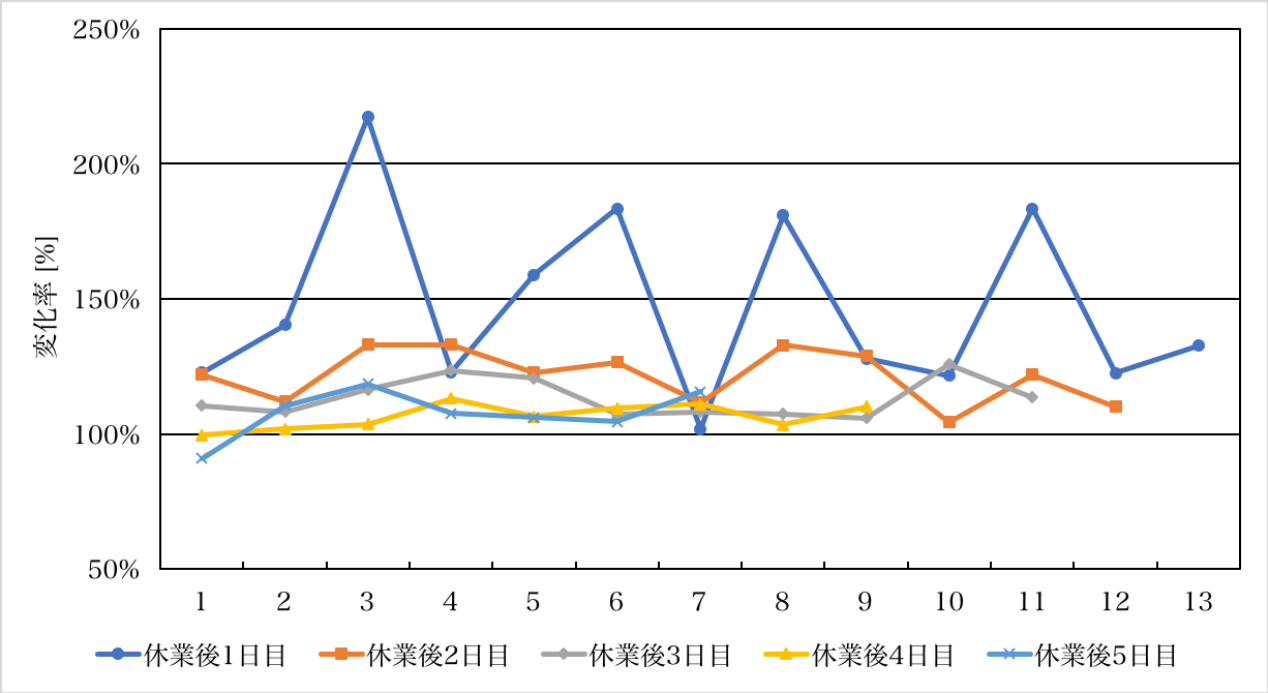

図 4.1　休業日後の経過日数による通常日に対する変化率の違い

表 4.3　休業日後の経過日数ごとの通常日に対する変化率の平均値と標準偏差の比較

	休業日後 1 日目	2 日目	3 日目	4 日目	5 日目
平均値	147.4%	121.5%	113.4%	106.5%	107.7%
標準偏差	32.7%	9.5%	6.8%	4.4%	8.3%

これらの結果より、予想通り休業日後の経過日数によって通常日に対する伸び率が異なることがわかる。したがって、もう一つの説明変数の候補を休業日後に対する重みとして、「休業日後 1 日目」、「休業日後 2 日目」、「休業日後 3 日目」、「休業日後 4 日目以降」という指標を追加する。一方、現状の説明変数の候補から、「受電率」を取り除く。その理由としては 2 つあり、1 つ目は求めるべき目的変数である着信件数が受電率の分母に入っているため、これを説明変数として利用するとモデルに影響を及ぼすのではないかと考えられるからである。2 つ目は受電率はあくまでも結果であり、事前に設定することができないため説明変数としてふさわしくないと考えられるためである（実際、X 社の方は受電率を使用する場合はあらかじめ 0.9 と設

定してモデルの説明変数として利用しているのが現状である)。また、計算機内では重回帰モデルの説明変数を「ダミー変数」から「ファクター (カテゴリ変数)」へ変更してモデルの構築を行う。

以上のことから、データ修正 B、新たな説明変数の追加および説明変数の型の変更を実施した上で、次節より 3 つのモデル案を提唱する。

4.2　モデル案 1

モデル案 1 として提唱するのは、曜日別 RIMA モデルを改良したモデルである。以下に、このモデルの説明を述べる。まずはじめに、現状の曜日ごとに分割した ARIMA モデルと同様に、データ修正 A を施した時系列データ $\{y'_t\}$ に関して曜日ごとに時系列を分割する。

$$\{y'_t\} = \bigcup_m \{y_{t_m,m}\} \qquad (\{y_{t_m,m}\}：m \text{ 曜日の時系列データ}) \tag{4.1}$$

その分割したそれぞれの時系列 $\{y_{t_m,m}\}$ に対して、「データの標準化」を行う。しかし、2015 年 6 月の水回りの修理に関する問い合わせの受付開始という環境の変化があるため、全体を標準化するのはよくない。その理由はその環境の変化の前後でデータ全体の水準（平均値）が異なるからである。そこで、2015 年 6 月前後でそれぞれ標準化を行い、組み合わせることで、その問題を回避する。曜日ごとに分割した時系列の 2015 年 6 月以前の部分 $\{y_{t_m,m,1}\}$ の平均値を $\mu_{m,1}$、標準偏差を $\sigma_{m,1}$、2015 年 6 月以後の部分 $\{y_{t_m,m,2}\}$ の平均値を $\mu_{m,2}$、標準偏差を $\sigma_{m,2}$ とすると、標準化を施した曜日ごとに分割した時系列 $\{\hat{y}_{t_m,m}\}$ は

$$\{\hat{y}_{t_m,m}\} = \frac{\{y_{t_m,m,1}\} - \mu_{m,1}}{\sigma_{m,1}} + \frac{\{y_{t_m,m,2}\} - \mu_{m,2}}{\sigma_{m,2}} \tag{4.2}$$

と表現できる。時系列データの標準化プロセスのイメージは以下のようになる。この標準化によって、課題 2 で取り挙げた図 3.7 や図 3.8 のような 2015 年 6 月時点での急激な変動を取り除くことができ、ARIMA モデルへの適合性がやや向上すると期待される。

この標準化を施した曜日ごとに分割した時系列 $\{\hat{y}_{t_m,m}\}$ に対して ARIMA モデルを適用し、曜日ごとに将来の予測を行う。

$$\hat{y}_{t_m,m}(1),\ \hat{y}_{t_m,m}(2),\ \cdots,\hat{y}_{t_m,m}(k) \quad \underset{\text{ARIMA}}{\Longrightarrow} \quad \hat{y}_{t_m,m}(k+1),\ \cdots,\hat{y}_{t_m,m}(k+i)$$

そして、その得られた予測値を $\mu_{m,2}$ と $\sigma_{m,2}$ を用いて再び元の形に戻す。

$$y_{t_m,m}(k+i) = \mu_{m,2} + \sigma_{m,2} \times \hat{y}_{t_m,m}(k+i) \tag{4.3}$$

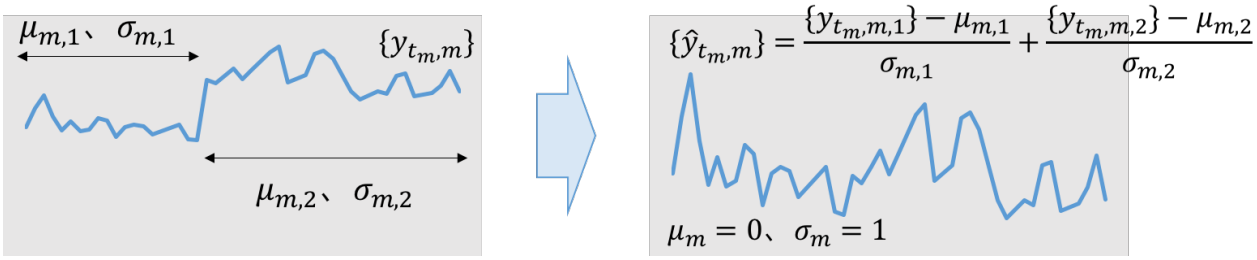

図 4.2　モデル案 1 におけるデータの標準化プロセスのイメージ

しかし、標準化を施すだけでは、課題 1 で挙げた祝日や 3 連休後などの特殊な日への対応はできていない。そこで、次は元の時系列 y_t とデータ修正 A を施した y'_t の差分を

$$z_t = y_t - y'_t \tag{4.4}$$

と定義すると、z_t は日付のカテゴリが「通常日」の場合は 0、そのほかの場合は 0 以外の数値となっている時系列である。つまり、z_t の 0 以外の数値は通常日に対する特殊な日の増減効果を表している。そこで、この特殊な日の増減効果 z_t を日付のカテゴリ毎に分割するが、通常日と休業日を除いた祝日 (前節で定めた連休中も含める)、3 連休後 (前説で定めた連休後も含める)、休業日前、休業日後 1 日目、休業日後 2 日目、休業日後 3 日目、休業日後 4 日目以降の 7 つのカテゴリの増減効果 z_l（ただし、l は前述の 7 つのカテゴリ）に着目する。そして、この増減効果 z_l に対してそれぞれ ARIMA モデルを適用し、増減効果の将来の予測を行う。

$$z_l(1),\ z_l(2),\ \cdots,z_l(h) \quad \underset{\text{ARIMA}}{\Longrightarrow} \quad z_l(h+1),\ \cdots,z_l(h+j)$$

以上より、データ修正 A を行い曜日ごとに分割した時系列にさらに標準化を施すことで求めた予測値 $y_{t_m,m}(k+i)$ に対して、必要に応じて特殊な日の増減効果の予測値 $z_l(h+1)$ を足し合わせることで、真の予測値を算出するような予測モデルを構築した。

4.3　モデル案 2

モデル案 2 として提唱するのは、階差をとることで季節性を取り除く ARIMA モデルを改良したモデルである。以下にこのモデルの説明を述べる。このモデル案 2 は、ARMA モデルに説明変数を加えたモデルである。モデルを式で表現すると、

$$\begin{aligned} y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \cdots \\ + \theta_q \epsilon_{t-q} + \beta_1 x_{1t} + \cdots + \beta_r x_{rt}, \quad \epsilon_t \sim N(0, \sigma_\epsilon^2) \end{aligned} \tag{4.5}$$

となる。ただし、定数項は省略する。これは、ARMA(p,q) のモデル (3.6) 式に説明変数 r 個 $(x_{1t},\ x_{2t},\ \cdots,\ x_{rt})$ から構成される回帰の項を付け加えたモデルである。よって、このモデルは一般的に ARMAX モデルと呼ばれる。ARMA(p,q) モデルと同様に、ラグ演算子 B を用いて表現すると、(3.8) 式より

$$\phi(B)y_t = \theta(B)\epsilon_t + \beta_1 x_{1t} + \cdots + \beta_r x_{rt}$$

$$\Longleftrightarrow \quad y_t = \frac{\theta(B)}{\phi(B)}\epsilon_t + \frac{\beta_1}{\phi(B)}x_{1t} + \cdots + \frac{\beta_r}{\phi(B)}x_{rt} \tag{4.6}$$

となる。ただし、$\phi(B)$ の零点の絶対値が 1 より大きい場合である。

そこで、このモデルへの適用方法を述べる。まず、使用するデータは元の時系列データから、「休業日」のデータを取り除いたデータを用いる。そして、この時系列データに対してデータ修正 B(日付のカテゴリの変更) を施した上で、上記の ARMAX モデルへ適用する。使用する説明変数は、表 3.2 の説明変数の候補に加えて、前節で新しく考えた「新規の修理に関する問い合わせの有無」と休業日後の重み付けとして「休業日後 1 日目」、「休業日後 2 日目」、「休業日後 3 日目」、「休業日後 4 日目以降」という指標を説明変数へと追加する。ただし、統計ソフト R でのモデル構築の都合上、このモデル案 2 においては説明変数の型をダミー変数とする。

以上より、当モデルに使用する時系列データと説明変数を上記のように設定し、予測する期間における説明変数の値を設定することで、現状の ARIMA モデルよりも精度が高い予測値が得られることが期待できる。さらに、課題 1 で取り挙げたような現状ではモデル外で行なっている特殊な日などの効果を経験的に補う作業の解消にも繋がると考えられる。

4.4　モデル案 3

モデル案 3 として提唱するのは、現状の重回帰モデルを改良したモデルである。以下に、このモデルの説明を述べる。課題 3 で説明したように、現状の重回帰モデルは誤差項に分散不均一性と系列相関が存在している。そこで、重回帰モデル (3.30) 式の誤差項 ϵ_t に ARMA(p,q) 過程を適用することで、それらの特徴を考慮したモデルへと変換を行う。誤差項 ϵ_t に ARMA(p,q) 過程を適用すると以下のように表される。

$$\hat{\epsilon}_{\mathrm{ARMA}} = \phi_1\epsilon_{t-1} + \cdots + \phi_p\epsilon_{t-p} + w_t + \theta_1 w_{t-1} + \cdots + \theta_q w_{t-q},\quad w_t \ \sim\ N(0,\sigma_w^2) \tag{4.7}$$

そして、この ARMA(p,q) 過程に従う誤差 $\hat{\epsilon}_{\mathrm{ARMA}}$ を重回帰モデルの目的変数 y_t から取り除く。

$$y_t' = y_t - \hat{\epsilon}_{\mathrm{ARMA}} \tag{4.8}$$

ARMA(p, q) 過程に従う誤差 $\hat{\epsilon}_{\mathrm{ARMA}}$ を考慮した y'_t を新たな目的変数とし、再び重回帰モデルを適用する。これによって、課題 3 への対応として誤差項の特徴を考慮したモデルとなり、予測モデルとしての適切さ (説明力) が向上することが期待される。

このモデルにおいても、モデル案 2 と同様の説明変数を使用する。ただし、今回の説明変数の型はファクター (カテゴリ変数) とする。目的変数も同じく元の時系列データから「休業日」を取り除き、データ修正 B(日付カテゴリの変更) を施したものを使用する。

以上より、重回帰モデルの誤差項に ARMA(p, q) 過程を適用し、かつ新たな説明変数を追加することによってモデルの適合度が良くなり、予測の精度も向上することが期待できる。また、誤差項に ARMA(p, q) 過程を適用することで、系列相関を取り除くことができ、分散不均一性についても緩和できることが期待できる。そこで実際に、モデル案 3 に対して Breusch-Pagan Test と Durbin-Watson test を実施すると以下のような結果となる。

```
> bptest(adj.reg.test.new)

        studentized Breusch-Pagan test

data:  adj.reg.test.new
BP = 297.29, df = 28, p-value < 2.2e-16

> dwtest(adj.reg.test.new)

        Durbin-Watson test

data:  adj.reg.test.new
DW = 2.0083, p-value = 0.4075
alternative hypothesis: true autocorrelation is greater than 0
```

図 4.3　モデル案 3 に対する Breusch-Pagan Test と Durbin-Watson test の結果

図 4.3 の結果より、重回帰モデルの誤差項の不均一分散に関しては p 値が 0.05 より小さいことから取り除くことができなかったが、系列相関については、DW 比が 2 に近いこと、p 値が 0.05 より大きいことから取り除くことができたと言える。

第 5 章

予測モデルの比較と考察

本章では、まず前章で構築したモデル案と現状のモデルとの予測精度の比較を行う。その内容を踏まえ、本研究で提唱する 3 つのモデル案に関して考察を行う。

5.1　予測の準備

モデルの予測精度の比較する上での準備を行う。まず、予測の期間の設定を行う。予測期間はデータが 2016 年 12 月 13 日まで存在しているので、2016 年 11 月まで実績値とモデルによる予測値の比較が可能である。よって、モデルの各月ごとの予測精度を確認するため、予測期間を直近 1 年間である 2015 年 12 月から 2016 年 11 月までと設定する。ただし、実務において予測の制約条件として、翌月 1 ヶ月の予測結果を当月の 15 日までに決定し、お客様相談センターに提供する必要がある。その例を図 5.1 に示す。

図 5.1　実務における予測の制約条件の例

そのため、実際的には当月の 12 日〜13 日までのデータを使用し、当月の残りの日数分と翌月 1 ヶ月分を予測しなければならない。したがって、本研究においてもこの制約条件に従い、約 2 週間の予測猶予期間と 1 ヶ月の求める予測期間を合わせて予測を行う。

次に、モデル案の予測精度を評価する指標としては以下の 2 つを使用する。

$$\text{RMSE (Root Meen Square Error：平均二乗誤差)} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \tag{5.1}$$

$$\text{MRE (Meen Relative Error：平均相対誤差)} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{y_i} \tag{5.2}$$

ただし、y_i は実績値、$\hat{y}_i$ は予測値、N は予測対象数である。RMSE は予測値が実績値からどの程度乖離しているかを示し、数値が 0 に近いほどモデルの予測精度が良いことを意味する。一方、MRE は X 社においてモデルの予測精度を測定する際に用いている指標でもあり、予測値と実績値の乖離が実績値に対して平均的にどの程度であるかを示している。この指標は数値が 0% に近いほどモデルの予測精度が良いことを意味している。

以上より、上記の予測期間およびその制約条件に従い、RMSE と MRE の 2 つの評価指標を用いてモデルの予測精度の比較を行う。

5.2　モデルの予測精度の比較および考察

予測精度の比較の対象とするモデルは、先行事例で述べた予測に使用されている、階差をとることで季節性を取り除いた ARIMA モデル、曜日ごとにデータを分割することで季節性を取り除いた ARIMA モデル、および重回帰モデルの 3 つのモデルと、本研究で提唱する 3 つのモデル案である。その比較結果をまとめたものが図 5.2 である。前述したように、各モデルに対して 2015 年 12 月から 2016 年 11 月までの直近 1 年間における 1 ヶ月ごとの予測を行っている（ただし、約 2 週間の予測期間も含めて予測を行なっている)。そして、各モデルの各月ごとの予測精度を RMSE と MRE の 2 つの指標で評価している。さらに、各モデルの各評価指標に対して年間の平均値、標準誤差、最小値（最も予測精度が高い結果)、最大値（最も予測精度が低い結果）をそれぞれ求めている。この比較表の見方としては、各行を見た際に最も優れた予測精度となっている箇所の背景が濃く、かつ数値が太く表示されており、それに準ずるような予測精度となっている箇所は背景のみ濃く表示されている。

そこで、この予測精度の比較表を確認し、各モデルについて考察を行う。毎月の予測精度の結果を見ると、現状のモデルの一つである重回帰モデルが直感的には良さそうに見える。しかし、年間を通してモデルの予測精度の良し悪しを判断するために、年間の平均値等を見ると、モデル案 2 が平均値も最も低く、かつ標準偏差も最も小さいことが読み取れる。さらに、最も予測精度が高い結果を示しているのもモデル案 2 であり、各モデルの最も予測精度が低い結果の中でも最も許容できる結果を示しているのもモデル案 2 である。したがって、モデル案

	現状のモデル						提案のモデル					
	階差ARIMAモデル		曜日ARIMAモデル		重回帰モデル		モデル案1		モデル案2		モデル案3	
	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE
2015年12月	346.6	16.0%	348.2	16.5%	**170.4**	**8.5%**	252.7	11.4%	176.0	9.6%	254.6	15.4%
2016年1月	387.2	20.2%	351.7	18.1%	339.0	21.3%	302.6	17.4%	**190.0**	**11.4%**	294.6	18.9%
2016年2月	277.5	16.8%	194.7	**11.7%**	**166.0**	**11.7%**	270.7	17.2%	201.2	14.4%	195.7	14.1%
2016年3月	321.8	17.2%	312.0	19.8%	**131.5**	**9.2%**	316.4	23.1%	218.7	17.3%	150.6	11.1%
2016年4月	177.9	8.8%	180.9	9.2%	106.1	7.1%	128.0	7.4%	**74.0**	**4.2%**	136.1	10.3%
2016年5月	315.7	9.9%	325.5	10.3%	219.8	**6.8%**	190.6	8.8%	**151.9**	7.7%	185.1	7.8%
2016年6月	161.3	8.8%	162.7	8.3%	186.0	12.8%	**114.7**	**6.5%**	137.6	9.1%	140.9	7.3%
2016年7月	403.5	22.1%	407.8	22.7%	**123.1**	**7.7%**	281.8	15.6%	242.6	11.4%	219.6	9.8%
2016年8月	194.3	6.4%	188.1	6.3%	171.6	10.1%	**119.4**	**5.5%**	139.7	7.0%	293.8	13.5%
2016年9月	411.3	18.6%	392.4	16.0%	**212.2**	**8.1%**	275.1	11.3%	232.6	8.6%	289.4	10.8%
2016年10月	325.4	12.1%	327.6	11.6%	268.8	10.4%	**204.9**	**8.3%**	230.2	9.3%	327.3	13.3%
2016年11月	487.4	28.3%	439.9	21.5%	**129.4**	**6.8%**	179.5	8.4%	219.1	14.2%	152.9	6.9%
平均値	317.5	15.4%	302.6	14.3%	185.3	**10.0%**	219.7	11.7%	**184.5**	10.4%	220.1	11.6%
標準誤差	96.4	6.2%	92.6	5.3%	64.2	3.9%	70.1	5.2%	**48.3**	3.5%	66.4	**3.4%**
最小値	161.3	6.4%	162.7	6.3%	106.1	6.8%	114.7	5.5%	**74.0**	**4.2%**	136.1	6.9%
最大値	487.4	28.3%	439.9	22.7%	339.0	21.3%	316.4	23.1%	**242.6**	**17.3%**	327.3	18.9%

図 5.2　各モデルの予測精度の比較表

	重回帰モデル		モデル案1		モデル案2		モデル案3		モデル案3'	
	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE	RMSE	MRE
2015年12月	**170.4**	**8.5%**	252.7	11.4%	176.0	9.6%	254.6	15.4%	215.7	12.4%
2016年1月	339.0	21.3%	302.6	17.4%	**190.0**	**11.4%**	294.6	18.9%	237.9	14.9%
2016年2月	**166.0**	**11.7%**	270.7	17.2%	201.2	14.4%	195.7	14.1%	185.9	13.3%
2016年3月	**131.5**	**9.2%**	316.4	23.1%	218.7	17.3%	150.6	11.1%	149.5	11.0%
2016年4月	106.1	7.1%	128.0	7.4%	**74.0**	**4.2%**	136.1	10.3%	98.8	6.3%
2016年5月	219.8	**6.8%**	190.6	8.8%	**151.9**	7.7%	185.1	7.8%	163.6	7.2%
2016年6月	186.0	12.8%	**114.7**	**6.5%**	137.6	9.1%	140.9	7.3%	136.2	8.9%
2016年7月	**123.1**	7.7%	281.8	15.6%	242.6	11.4%	219.6	9.8%	163.1	**6.9%**
2016年8月	171.6	10.1%	**119.4**	**5.5%**	139.7	7.0%	293.8	13.5%	141.2	7.7%
2016年9月	**212.2**	**8.1%**	275.1	11.3%	232.6	8.6%	289.4	10.8%	242.6	9.1%
2016年10月	268.8	10.4%	**204.9**	**8.3%**	230.2	9.3%	327.3	13.3%	310.5	12.5%
2016年11月	**129.4**	**6.8%**	179.5	8.4%	219.1	14.2%	152.9	6.9%	143.2	**6.7%**
平均値	185.3	10.0%	219.7	11.7%	184.5	10.4%	220.1	11.6%	**182.4**	**9.7%**
標準誤差	64.2	3.9%	70.1	5.2%	**48.3**	3.5%	66.4	3.4%	56.6	**2.8%**
最小値	106.1	6.8%	114.7	5.5%	**74.0**	**4.2%**	136.1	6.9%	98.8	6.3%
最大値	339.0	21.3%	316.4	23.1%	**242.6**	17.3%	327.3	18.9%	310.5	**14.9%**

図 5.3　追加検証による予測精度の比較表

2 は他のモデルに比べて、予測精度が高くかつ安定したモデルであるということができる。一方、現状の重回帰モデルに関しては、予測精度の平均値は高いものの、標準偏差が高くなっている。この原因は、2016 年 1 月の予測精度が低くなってしまっているためである。この月の予測精度が悪くなった理由としては、ちょうど 2015 年から 2016 年にわたる予測であり、モデルの説明変数の中に 2016 年の情報がないため、年を説明変数から除外してモデルを構築しなければならず、求める着信件数のうち年によって説明できる部分をうまくモデルで予測できていないためではないかと考えられる（2016 年の情報がない状態で年を説明変数として使用して予測を行なった場合は、より悪い予測精度となる）。その観点から比較すると、モデル案 1 に関してはもとより年による効果は考慮しておらず、またモデル案 3 に関しては重回帰モデルと同様に年を説明変数から除外せざるを得ないが、モデル案 2 に関しては、このモデルにおいても説明変数として年を使用できないが、その分を ARMA モデルの特徴である近い過去の自分自身の回帰によって、本来年によって説明できる部分を補えているのではないかと推測できる。そのため、モデル案 2 は 2016 年 1 月は他のモデルに比べて予測精度が高くなっていると考えられる。

また、モデル構築をする際に 4 つの課題を設定したが、これらの課題が解決したのかどうかを予測精度の結果をもとに検証を行う。まず、課題 1 に関しては、現状の ARIMA モデルはモデル外で経験的に祝日や 3 連休後などの特殊な日の効果を補っているため、モデル案 1 とモデル案 2 によってこれらをモデル内で取り扱えるようになったことは改善点である。実際に、図 5.2 からも現状の ARIMA モデルよりもモデル案 1 とモデル案 2 の方が予測精度が高くなっている。ARIMA モデル外で経験的に補う手法に関して厳密にはわからないが、MRE で約 10% 程度であるそうなので、悪くはない予測精度であると考えられる。次に課題 2 に関しては、標準化を行うことで、新規の問い合わせという変化によって定常性が失われることを回避しようと試みたが、モデル案 1 の実際の予測精度に関しては、2016 年 6 月（祝日や 3 連休後などの特殊な日が一日も存在しない月）において、現状の曜日 ARIMA モデルと比較すると、予測精度はモデル案 1 の方が高く、データを標準化することによって多少の改善は得られたのではないかと考えられる。課題 3 に関しては、重回帰モデルの誤差項の分散不均一性と系列相関のうち、系列相関を考慮したモデルを考えたが、予測精度の観点からすると、現状の重回帰モデルと新規のモデル案 3 を比較してわかるように、期待される改善は得られなかったと言える。そこで、追加検証を行う。モデル案 3 は説明変数の型を「ファクター」として計算を行なっているが、現状の重回帰モデルと同様に説明変数の型を「ダミー変数」として計算を行う。このモデルをモデル案 3’ とし、現状の重回帰モデルおよび本研究で提唱する 3 つのモデル案との予測精度の比較結果をまとめたものが、図 5.3 である。この追加検証による予測精度の比較表からわかるように、モデル案 3’ は各月の予測精度は他の予測モデルに多少劣ってはいるものの、年間の平均値や標準偏差等を見ると、他のモデルと比較して優れた予測精度を示

している。したがって、モデル案 3' はモデル案 3 の説明変数の型を変更して計算を行うことで、その予測精度は現状の重回帰モデルから若干の改善を得ることができたと考えられる。最後に課題 4 に関しては、図 5.2 からは直接的な改善効果を確認することはできないが、説明変数の追加や日付カテゴリの変更も予測精度の向上につながっていると考えられる。

次に、各予測モデルによる予測値と実績値の推移を示すと、図 5.4〜図 5.7 のようになる。実績値は破線 (青)、予測値は実線 (橙) で表されている。そこで各モデル案について考察を行う。まずモデル案 1 については、図 5.4 を見ると 2016 年 1 月から 3 月にかけては過大予測となっており、一方、7 月から 10 月にかけては過小予測となっている。モデル案 2 については、図 5.5 を見ると全体的に極小点は多く、極大点は少なく予測される傾向があることがわかる。モデル案 3(3') については、モデル案 1 と同様に 1 月から 3 月にかけては過大予測となっており、7 月から 10 月にかけては過小予測となっている。モデル案 1 に関しては、月による違いを考慮していないため月によって多く予測したり少なく予測したりしてしまう可能性があるのは理解できるが、モデル案 2 および 3(3') に関しては、説明変数として「月」を使用しており、月による違いを考慮しているため、本来ならば 1 年間のマクロなトレンドをある程度捉えることが期待できるはずである。しかし、実際には 1 年間のマクロなトレンドを正確に予測できていない。したがって、月よる違いをうまくモデルに取り入れることができていないのか、もしくは月による違いとは別の要因が部分的に存在するのか、と考えることができる。前者の場合であれば、例えばモデル外でさらに年間のマクロなトレンド (近似的に線形なトレンド) を追加することで、さらなる予測精度の向上が期待できる。後者の場合であれば、「トイレ」という商品に関して、問い合わせが増減するような要因を X 社の方と協議し、そのような要因（特に、7 月から 10 月にかけてトイレに関しての問い合わせが増加するような要因）があれば説明変数としてモデルに組み込むことで、さらなる予測精度の改善が期待できる。

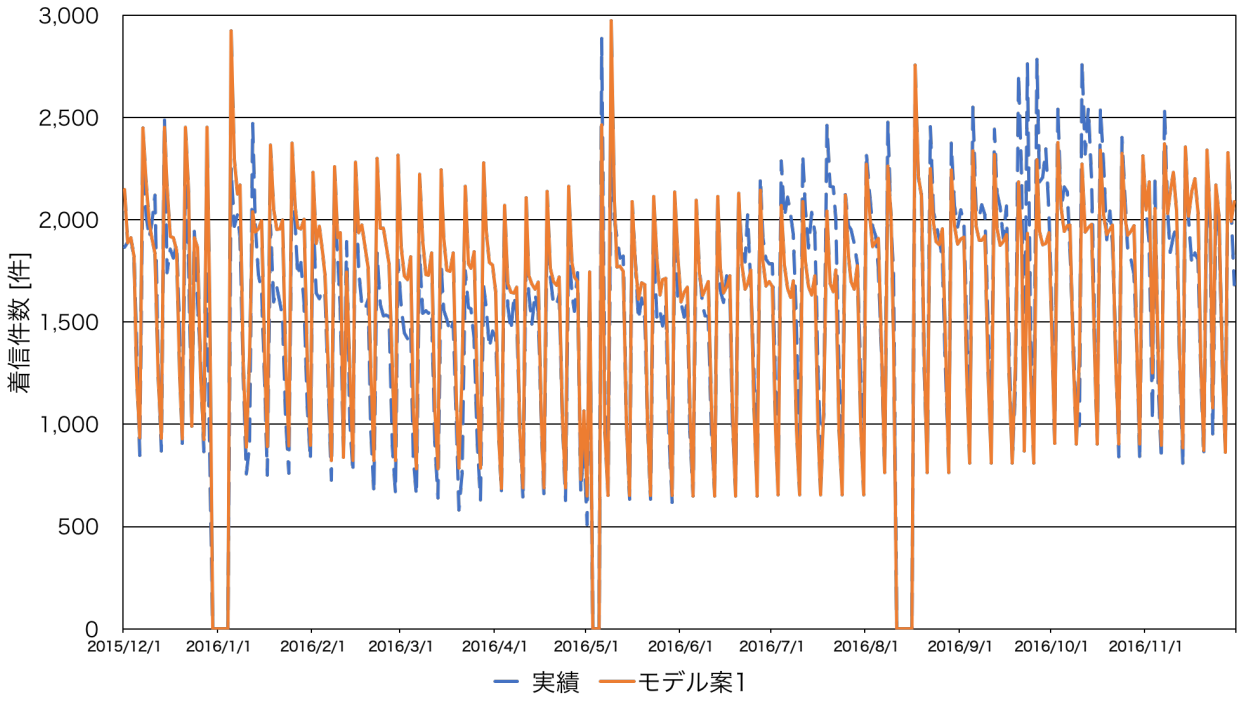

図 5.4　モデル案 1 による予測値と実績値の比較

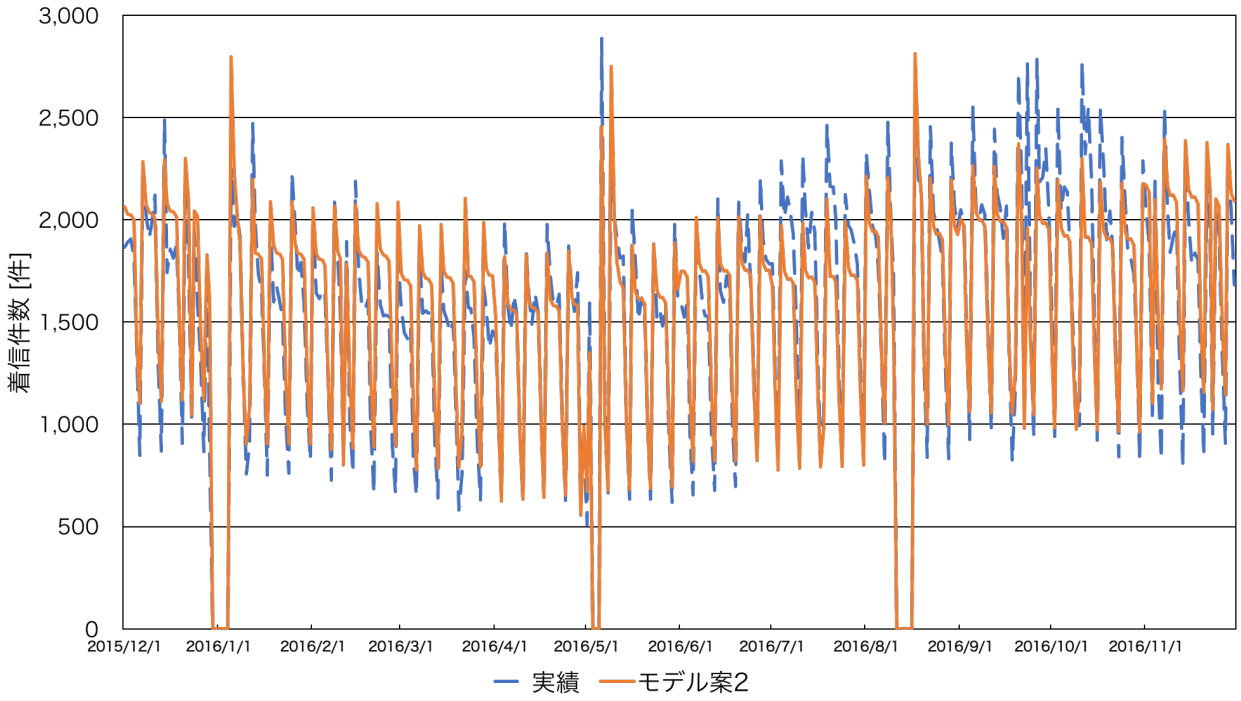

図 5.5　モデル案 2 による予測値と実績値の比較

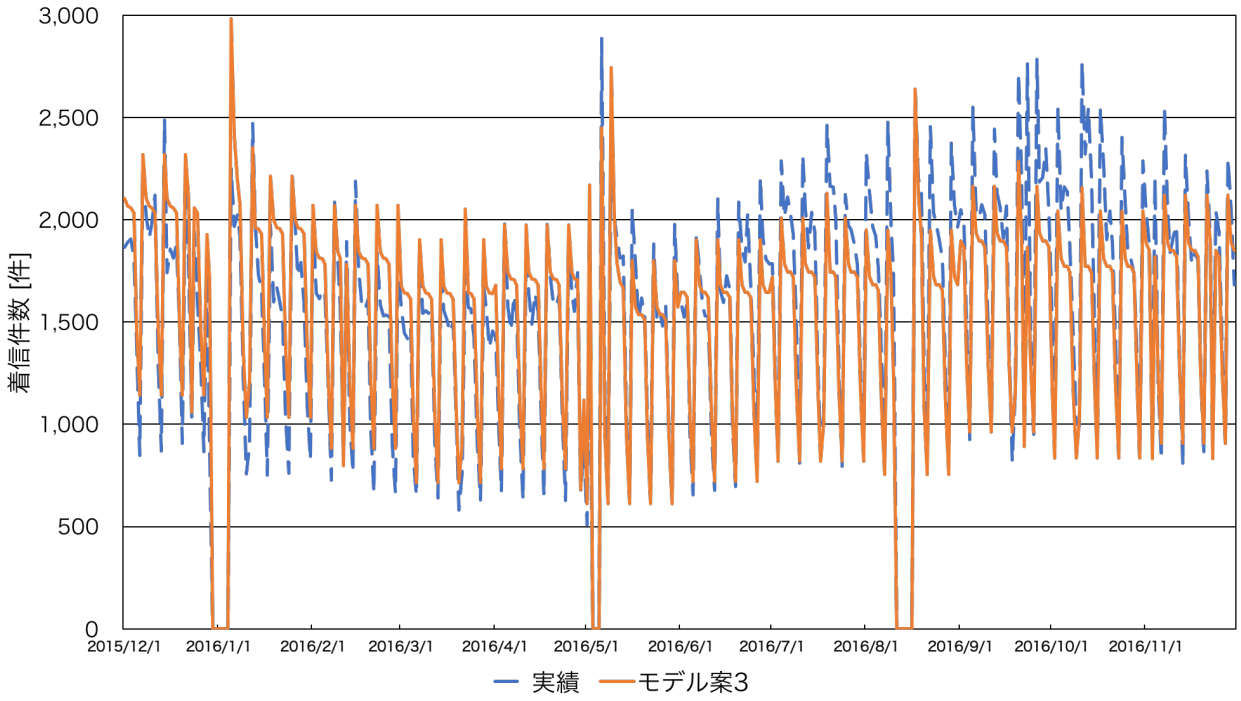

図 5.6　モデル案 3 による予測値と実績値の比較

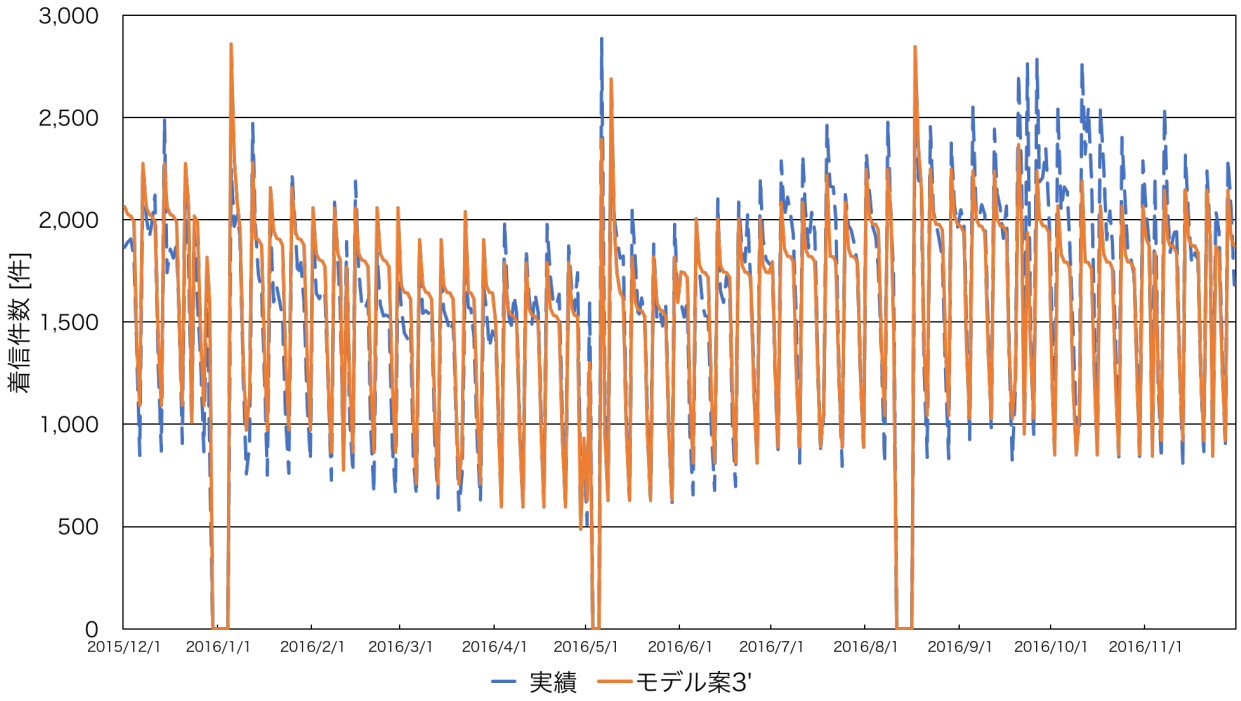

図 5.7　モデル案 3' による予測値と実績値の比較

第 6 章

結論

本研究では、X 社のお客様相談センターの着信件数の予測モデルの構築を行った。現状の予測モデルよりも予測精度が高いモデルを構築するという目的に対して、3 つの予測モデル案を提案した。詳しく説明すると、まず現状のモデルおよび使用するデータの特徴を把握し、4 つの課題を設定した。1 つ目は現状の予測に使用されている ARIMA モデルでは日付のカテゴリの特徴が考慮されていないことである。2 つ目は現状の予測に使用されている ARIMA モデルでは新規の問い合わせによって時系列データの定常性が失われており、差分をとったりデータを分割したりするだけでは、その時点においての急激な変化を取り除けないということである。3 つ目は重回帰モデルの誤差項が制約条件を満たしていないということである。4 つ目は説明変数の追加の検討と見直しである。これらの課題を踏まえモデルを構築する前に、データの前処理として、4 つ目の課題について、水回りの修理に関する受付の有無と休業日後に対して経過日数を考慮して重み付けをする説明変数を追加し、一方、受電率を説明変数の候補から取り除き、加えて日付カテゴリの見直しを行うことで解決した。そこで、残りの 3 つの課題を解消するために、3 つの改良型モデルを提案し、それらのモデルの予測精度の比較を行った。モデル案 1 では曜日ごとに分割したデータを標準化し、ARIMA モデルに適用し、さらに日付カテゴリによる増減効果を算出し、それらに対しても ARIMA モデルを適用し、最終的に各予測結果を組み合わせるというモデルである。このモデル案 1 は課題 1 および課題 2 の解消を目指したモデル案であり、予測精度の観点からは多少の改善は得られたと考えられる。一方、モデルの煩雑性の改善にまでは至らなかった。モデル案 2 では、現状の ARIMA モデルに対して、説明変数を加えるモデルである。予測精度の観点では、現状のモデル、そして本研究で提唱する 3 つのモデル案の中で最も優れた予測モデルであると考えられる。このモデル案も課題 1 を解消に導いたと言える。最後にモデル案 3(3') は重回帰モデルの誤差項に ARMA(p, q) 過程を適用したモデルであり、課題 3 のうちの系列相関の除去に成功した。同様に予測精度の観点では、説明変数の型を「ダミー変数」にしたモデル案 3' はモデル案 2 と同等の予測精度であり、現状の重回帰モデルから改善することができたと言える。しかし、誤差項の分散不均

一性は除去することができていないためさらなる改善の余地はある。そして、3 つのモデル案に共通して言えるのは、1 年間のマクロなトレンドをモデルにうまく反映しきれていないということである。

今後の展望をまとめる。まず予想精度に関しては、本研究で提唱するモデル案で再現しきれていない部分的な予測の過不足への対応である。その原因が月による違いをモデルに取り入れることができていないのであれば、モデルに春から夏にかけて上昇し、秋から冬にかけて下降するような近似的な線形トレンド項を補うことで改善が期待できる。一方、その原因が月による違いとは別の要因にあるのであれば、今後 X 社の方とともに協議し、新規の CM を始めたとか新製品をリリースしたとかなどといったトイレの着信件数を増減させるような要因を特定し、モデルに組み込むことで改善が期待できる。また予測精度以外に関しては、例えばモデルの選択基準を構築することが考えられる。これに関しては、X 社にとって予測を過大にしてしまうことと過小にしてしまうことのどちらがより経営課題として重大であるかを考慮する必要がある。さらに、予測をどちらの方向にどの程度外すことで X 社のお客様相談センターにどの程度のインパクトを及ぼすのかという情報が得られるのであれば、予測精度だけではなく、様々な観点から「良い予測モデル」を判断できるのではないかと期待できる。そうすることで、お客様相談センターのオペレーターの効率的な調達、育成および配置が実現可能となるであろう。また、最終的に予測モデル選択の自動化を目指すためには、直近の最も予測精度が高いモデルを逐次選択していくようなモデルを構築するべきではないかと考えられる。

最後に、本研究で提案した 3 つのモデル案の予測精度について検証してきたが、月ごとに最も優れた予測精度を示すモデルは変わっており、どのような年月の予測にも対応できるような万能なモデルの構築は難解であることを再認識した。しかし、今後も可能な限り万能なモデルに近づけるように努めていきたいと考えている。

参考文献

[1] 坂巻隆治 *et al.*, データサイエンティスト養成読本 R 活用編, 技術評論社, (2015)
[2] 北川源四郎, 時系列解析入門, 岩波書店, (2005)
[3] P. J. ブロックウェル / R. A. デービス, 入門 時系列解析と予測, シーエーピー出版, (2000)
[4] 田中孝文, R による時系列分析入門, シーエーピー出版, (2008)
[5] 永田靖 / 棟近雅彦, ライブラリ新数学大系 = E20 多変量解析法入門, サイエンス社, (2001)
[6] 金明哲, R によるデータサイエンス ■データ解析の基礎から最新手法まで■, 森北出版, (2007)
[7] 福地純一郎 / 伊藤有希, R による計量経済分析【シリーズ】統計科学のプラクティス, 朝倉書店, (2011)

謝辞

本研究を進めるにあたり、終始親切なる御指導と御鞭撻を賜り、本論文をまとめるに際しても、親身な御助言と力強い励ましを頂いた、修士論文指導教員の慶應義塾大学大学院経営管理研究科 林高樹教授に、深く御礼を申し上げます。また、快く副査を引き受けてくださった慶應義塾大学大学院経営管理研究科 高橋大志教授、市来嵜治専任講師、住田潮特任教授にも、深く感謝を申し上げます。そして、本研究を共同研究としてデータの御提供及び御助言を頂いた、X 社の担当者の方々に深く感謝を申し上げます。また、研究員として本年度のゼミ生を見守ってくださった M37 の森さん、研究内容は違えど互いに高め合ってきたゼミ同期の吉岡さん、上田さん、守谷君、内山君、常松君、皆のおかげでこの 1 年間のゼミ生活を有意義な時間にすることができました。心より感謝致します。M38 の同期の皆さんも人生の先輩として様々なアドバイスやサポートをしていただき、感謝致します。

最後に、これまで私を支えてくださった多くの方々、そして色々な面で学生生活を支えてくれた家族や友人に感謝致します。