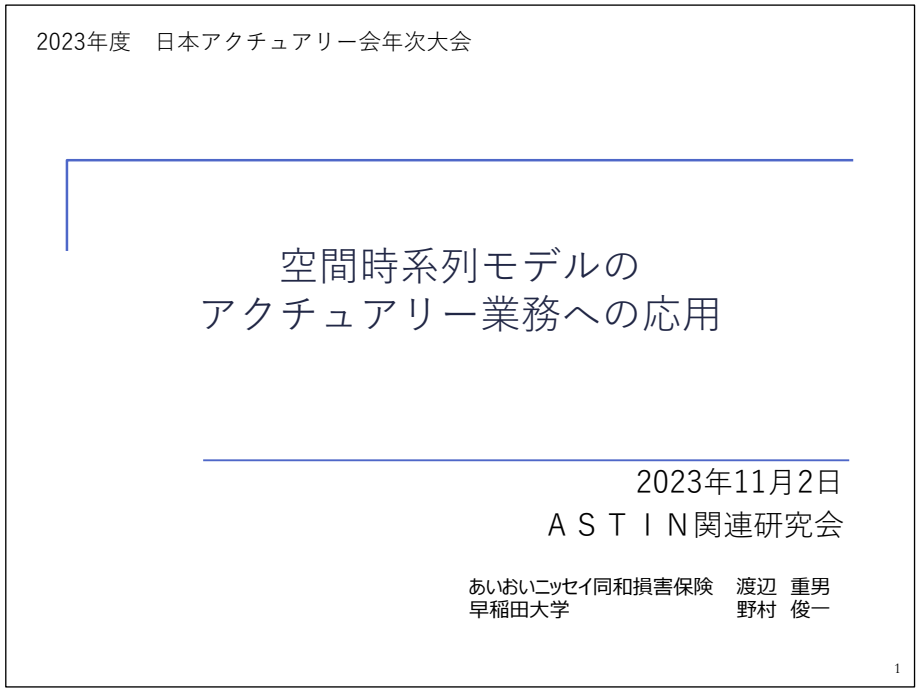
空間時系列モデルのアクチュアリー業務への応用
＜ASTIN 関連研究会＞

早稲田大学　野村　俊一　君

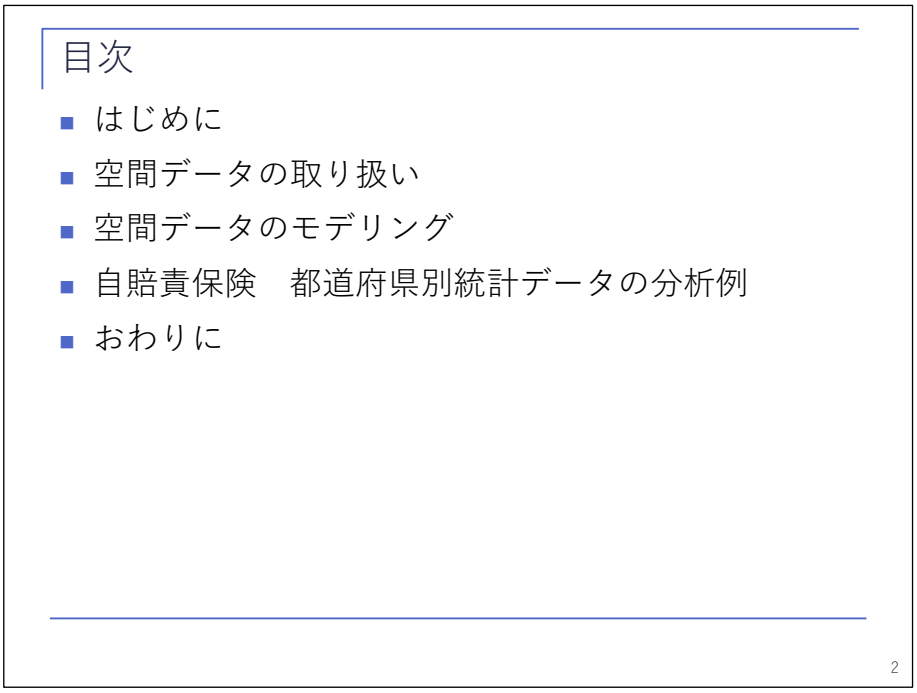
あいおいニッセイ同和損害　渡辺　重男　君

【司会】　時間となりましたので、セッション A-5、ASTIN 関連研究会による「空間時系列モデルのアクチュアリー業務への応用」を開始します。発表者は、あいおいニッセイ同和の渡辺重男さん、早稲田大学の野村俊一さんのお二人ですが、都合により、野村さんが本日会場に来られなくなりましたので、野村さん発表分については、事前に録画した動画を放映させていただきます。

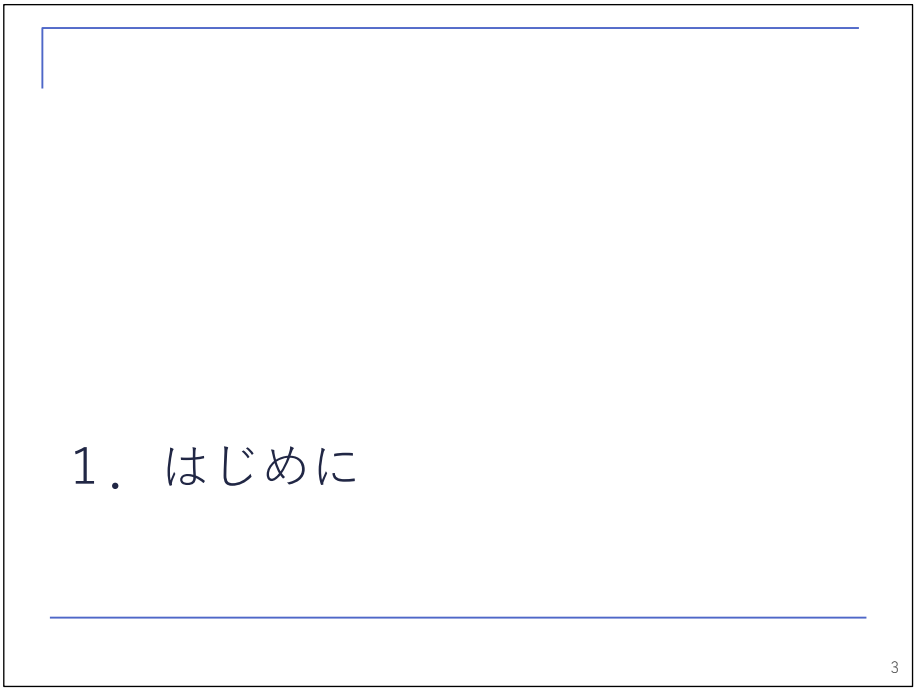
なお、質疑応答の時間は、お二人が発表したあとにまとめて取らせていただきます。また、Slido に投稿された質問に対しても、その際に回答することといたします。オンラインで視聴されている方で、質問のある方は、Slido への投稿をお願いします。それでは、渡辺さん、よろしくお願いします。



【渡辺】　ASTIN 関連研の渡辺です。よろしくお願いします。このセッションでは、「空間時系列モデルのアクチュアリー業務への応用」というタイトルで、ASTIN 関連研の活動として輪読を行った書籍の内容をベースに、お話ししたいと思います。



本日の発表の内容は、こちらのスライドのとおりです。はじめに輪読を行った書籍について簡単に紹介しまして、次の二つのパートで、本の内容をベースに、空間データの取り扱いと空間データのモデリングについてお話しします。そのあと、実際の分析例として、自賠責保険の都道府県別統計データを使った例をご紹介します。前半は私からお話をしまして、後半は早稲田大学の野村さんからお話をします。ただ、野村さんは、冒頭にお話がありましたように、都合により事前に用意した動画での発表となりますのでご了承ください。



本セッションの概要

- 一般に、保険事故の発生率にはある程度の地域差があり、それは時間的にも変化している。そのような時空間データを分析するには、時間的・空間的な相関関係を取り入れたベイズモデルが有用となる
- 本発表では、健康・医療データに対する時空間ベイズモデルを解説した書籍「Using R for BayesianSpatial and Spatio-Temporal Health Modeling」に基づき、空間データの扱いと R による時空間ベイズモデルの実装について解説する
- さらに、時空間ベイズモデルを自賠責保険の都道府県別統計データに適用した以下 2 つの分析例を紹介する
  - 新型コロナウイルス感染流行期における事故発生率変化の分析
  - クレーム頻度・単価に基づく都道府県クラスタリング分析

4

では、本題に入ります。本セッションの概要は大会プログラムに記載されているとおりですので飛ばします。

書籍の紹介

- Lawson, A. B. (2021). Using R for Bayesian spatial and spatio-temporal health modeling. CRC Press. https://www.routledge.com/Using-R-for-Bayesian-Spatial-and-Spatio-Temporal-Health-Modeling/Lawson/p/book/9780367760670
  - ベイズ階層モデルによる空間的健康データの分析（Bayesian Disease Mapping、ベイズ疾患マッピング）の入門書
  - BRugs/OpenBUGS, CARBayes, Nimble, INLAによる実装を紹介し比較
  - 著者サイト https://people.musc.edu/~abl6/ 本書で使用されているデータやコードなど

5

続いて本の紹介です。本セッションで取り上げる本は、ベイズ階層モデルによる空間的健康データの分析、すなわちベイズ疾患マッピングの入門書でして、R による実装について、RBugs や CARBayes、Nimble、INLA などいろいろなツールでの実例を紹介している点が特徴です。この本で紹介されているデータやコードなどは、著者のサイトで公開されていますので実際に手を動かしながら使い方を学んでいただくことができるかと思います。

書籍の紹介 (続)

	Chapter 1	Introduction and Data Sets
	Chapter 2	**R Graphics and Spatial Health Data**
	Chapter 3	Bayesian Hierarchical Models
	Chapter 4	Computation
	Chapter 5	Bayesian model Goodness of Fit Criteria
	Chapter 6	Bayesian Disease Mapping Models
Part I Basic Software Approaches	Chapter 7	BRugs/OpenBUGS
	Chapter 8	Nimble
	Chapter 9	CARBayes
	Chapter 10	**INLA and R-INLA**
	Chapter 11	Clustering, Latent Variable and Mixture Modeling
	Chapter 12	Spatio-Temporal Modeling with MCMC
	Chapter 13	**Spatio-Temporal Modeling with INLA**
Part II Some Advanced and Special topics	Chapter 14	Multivariate Models
	Chapter 15	Survival Modeling
	Chapter 16	Missingness, Measurement Error and Variable Selection
	Chapter 17	Individual Event Modeling
	Chapter 18	Infectious Disease Modeling

6

こちらが、この本の目次になります。色をつけてある所が、今回の発表に特に関係が深い部分ですが、それ以外で言いますと、特に最後の五つの章が、より進んだ内容、あるいは個別のトピックを扱った所です。中でも最後の 18 章では、感染症のモデリングを扱っておりまして、最近ですと新型コロナの感染拡大モデルの基礎になった SIR モデルについて解説されています。

ベイズ疾患マッピングを学ぶ意義

- リスクの空間的・時間的な振る舞いをモデル化するベイズ疾患マッピングは、疫学・公衆衛生学等の分野において重要な役割を果たしている
- アクチュアリーが扱うデータには、地理情報を持つ時系列データとみなせるものが多い
- ベイズ疾患マッピングの手法をアクチュアリー業務に応用することで、より予測性能が高く、また説明しやすいモデルを構築することができるのではないか

7

本書で扱われているベイズ疾患マッピングは、リスクの空間的・時間的なふるまいをモデル化する手法として、疫学・公衆衛生学等の分野で重要な役割を果たしています。このようなモデルをアクチュアリー業務に応用することで、よりよいモデルを構築できるようになるのではないか。そのような点にベイズ疾患マッピングを学ぶ意義があるかと思います。

公募型共同研究としての取り組み

- 情報・システム研究機構 データサイエンス共同利用基盤施設の公募型共同研究(https://ds.rois.ac.jp/crp/)の採択課題「保険数理データ解析のための現代的数理手法の開発」の一環として、ASTIN関連研究会と大学の有志で本書の輪読を実施した
- さらに、本書の手法を用いた以下の研究（括弧内は研究メンバー）を進めており、本発表の後半で紹介する
  - 新型コロナウイルスの感染流行期における都道府県別自動車事故の分析（谷川祥太、野村俊一、佐野誠一郎）
  - 損害保険料率算定における空間ベイズモデルを用いた空間クラスタリング（大竹隼登、野村俊一）

8

今回ご紹介する本の輪読は ASTIN 関連研で取り組んだものではありますが、一方でここにありますような、情報・システム研究機構データサイエンス共同利用基盤施設の公募型共同研究という側面もあります。このような活動を通じて、実務家と大学の先生、学生さんも含めて、一つのテーマについて意見交換をしながら進めるという非常によい機会になっていると考えています。

過去の発表との関係

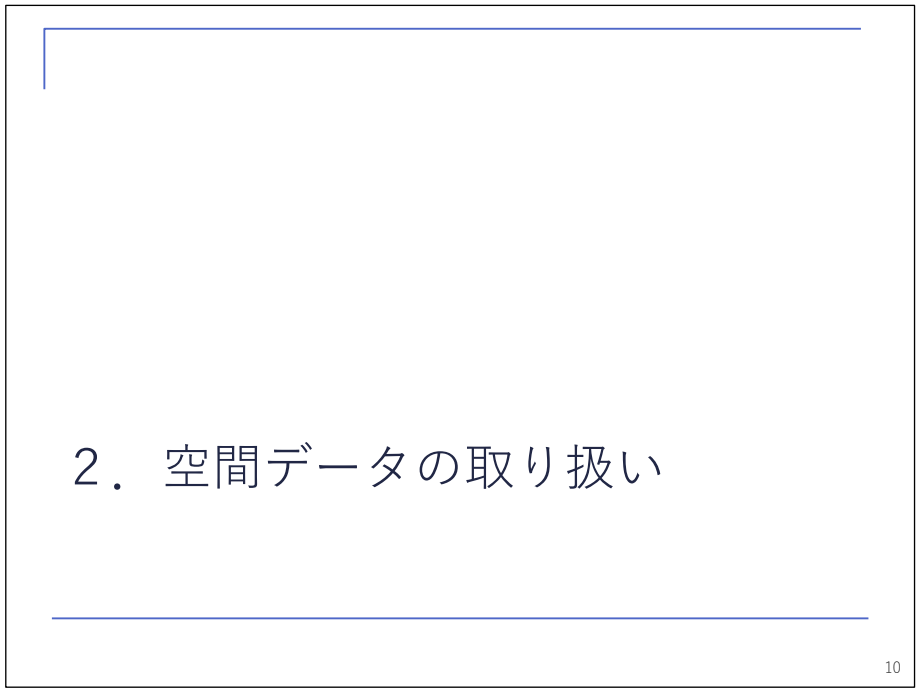
- 空間データの取り扱いの詳細、ベイズ的手法全般やMCMCによる実装については、2022年度年次大会のプレゼンテーション参照
  - 「Rによる計算保険数理」

    HOME ＞ ライブラリ ＞ 年次大会報告集 ＞2022年度 年次大会報告集

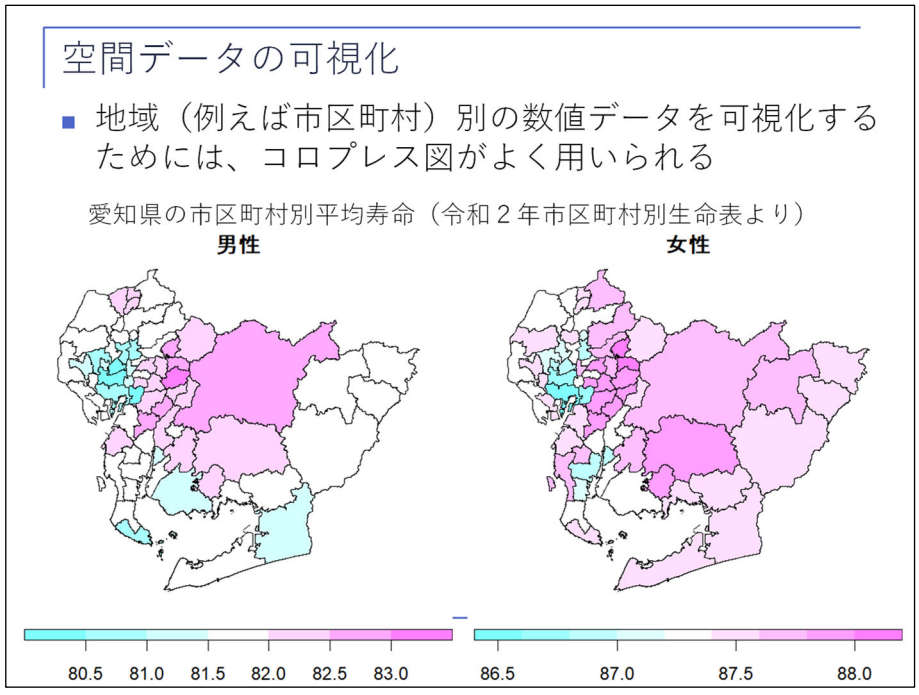
9

「はじめに」の最後ですが、昨年も ASTIN 関連研でプレゼンテーションを聞いていただいた方は、同じようなテーマをやっていると思われかもしれません。昨年度の年次大会で「R による計算保険数理」というタイトルでプレゼンテーションを行っていまして、その中でも空間データやベイズについて扱っています。ただ、昨年度は、空間データの取り扱いの細かいところやベイズの全般的な話、また、MCMC の実装に

ついてお話ししております。今回のプレゼンテーションは、なるべくそれとかぶらないようにしています。逆に前年度紹介した内容については触れていませんので、適宜そちらの方を参照いただくようお願いします。



ここからが本題です。まず、空間データの取り扱いついてお話ししたいと思います。



最初に、空間データの可視化を考えます。市町村別などの地域別統計データ、数値データを可視化するための手法として、よくこのような図が使われているかと思います。このような図を作ってみようと考えます。

地図情報

- 空間データの可視化やモデリングにあたり、まずは地図情報を用意する必要がある
- 地図情報を扱う際には、地図上の領域の頂点の情報を持つPolygon（ポリゴン、多角形）データがよく用いられる
- ポリゴンデータは、一般にシェイプファイルの形式で提供される
  - シェイプファイルは、拡張子shp, dbf, shx, prjを持つファイルの集合体
  - 詳細は2022年度年次大会プレゼンテーション参照

12

そのためには、まず地図情報を用意する必要があります。地図情報を扱う際には、地図上の、市区町村などの領域の頂点の情報を持つポリゴンデータがよく用いられます。このポリゴンデータは自分で作ることもでき、前年度のプレゼンテーションではその方法を細かくお話ししたと思います。

一方で、できあいのシェイプファイル等の形式で提供される既存データも利用できますので、そちらを使うこともよくあります。

地図情報②

- 国土交通省「国土数値情報」
  - 国土に関する基礎的な情報をGISデータとして整備し提供
  - 利用の際は出典の記載が必要
    「出典：国土交通省国土数値情報ダウンロードサイト（https://nlftp.mlit.go.jp/ksj/index.html）」
  - 「行政区域データ」から必要な地域を選択しダウンロード
  - シェイプファイル（拡張子shp, dbf, shx, prj ）を適当なフォルダにコピーし、sfパッケージのst_read()で読み込む
    （文字コードがUTF-8なので、読み込む際に注意が必要）

```
library(sf)
map_jp <- st_read(
  "<保存先フォルダ名>/N03-23_230101.shp",
  options="ENCODING=CP932")
```

13

例えば国土交通省は、国土に関する基礎的なデータを整理して「国土数値情報」として提供しています。その中には行政区域データもありまして、スライドにある青字のコードのようにして R に読み込むことができます。今回は、これを使っていきたいと思います。

スライドの 1 行目で、sf パッケージというものを使って読み込んでいます。また、コードの最後の行では、options という引数で文字コードを指定しています。この国土数値情報のデータは、Windows で使っている文字コードと違うものを使っていますので、このオプションにより Windows で使っている文字コードである CP932 に変換して読み込むということを指定しています。

読み込んだデータはどのようなものかというと、このよう形式になっています。いろいろな情報が含まれていますが、特に赤字の所に注目していただくと、一番上で、データの形式としては sf パッケージで使われるデータ形式である sf オブジェクトであり、同時に data.frame でもあることが分かります。

data.frame の列は 9 個の変数が含まれており、中身はここにあるとおりです。一番上が、1 から始まって 12 万ぐらいまである連番ですが、その下の四つの項目が、見て分かるとおり、行政区画の名前が入っています。その下に 5 桁のコードが入っていますが、これは行政区画コードで各市区町村に振られているコードです。その下の二つが、それぞれのポリゴンの周の長さと面積です。単位はキロメートルとかということではなくて、そのまま使うわけにはいかないのでご注意ください。一番下に geometry という項がありますが、これは、それぞれのポリゴンの頂点の情報を含むデータになっています。

このデータから冒頭の図を作ろうと思うと、愛知県の情報を抜き出してこなければいけません。

地図情報④

- 国土交通省「国土数値情報」（続き）
  - うち愛知県（都道府県コード23）のデータを抽出 (a)
    - この状態では、離島や埋め立て地等は別レコード
  - 各市区町村に対応するレコードがユニークになるよう統合(b)
  - 所属未定地「名古屋港口埋立地」（コード23801）を除外 (c)
    - (b)(c)の操作をすれば市区町村別データとの紐づけが可能

```
(a) map_jp.aichi <- map_jp[grep("23...", map_jp$N03_007), ]

(b) map_jp.aichi <- aggregate(map_jp.aichi,
      by=list(map_jp.aichi$N03_007), FUN=unique)

(c) map_jp.aichi <-
      map_jp.aichi[map_jp.aichi$N03_007!="23801", ]
```

15

そこで、まず(a)のコードを使って、行政区画コードが 23 で始まるのが愛知県ですので、これを抜いてきます。ただ、抜いてきた状態だと、それぞれポリゴンごとに一つのレコードが作られていますので、島や埋立地があったりすると別のレコードになってしまって、市区町村のデータと紐づけができませんので、(b)のコードを使って市区町村ごとに一意のデータに統合していきます。ここまでやれば使えはするのですが、このまま使うと、市区町村に紐づいていないデータがあると、色が塗られずに常に白くなって見栄えが悪いので、最後は(c)のところでこれも除外しています。

地図情報⑤

- 国土交通省「国土数値情報」（続き）

```
plot(map_jp.aichi[, 1], col=NA, main="")
```

geometry（頂点の座標等）を除き8つの列が含まれており、指定しなければすべての列についてプロットが作成される

16

このように作ったデータをプロットすると、このスライドのとおりになります。ここに、これから先、数値情報を乗せていこうと思います。

データの可視化

- コロプレス図（Choropleth map）
  - 地図上の領域別（例えば市区町村別）の統計情報を、色の濃淡や異なる色調を使って表現
  - sfオブジェクトに必要なデータを含む列を追加し、plot()でプロット
  - 例として、市区町村別の平均寿命をプロット
    - 元データ：令和２年市区町村別生命表(*1)
    - 愛知県の各市区町村の男女別平均寿命を抽出したCSVファイルを作成（並び順はsfオブジェクトにあわせコード順とする）

コード	都道府県	市区町村	平均寿命.男性	平均寿命.女性
23101	愛知県	名古屋市千種区	82.2	88
23102	愛知県	名古屋市東区	81.7	87.4
23103	愛知県	名古屋市北区	80.9	86.9
23104	愛知県	名古屋市西区	80.9	87.4
23105	愛知県	名古屋市中村区	80.3	87

*1 https://www.e-stat.go.jp/stat-search/files?page=1&layout=datalist&toukei=00450012&tstat=000001031336&cycle=7&tclass1=000001060926&tclass2=000001204500&cycle_facet=tclass1&tclass3val=0

17

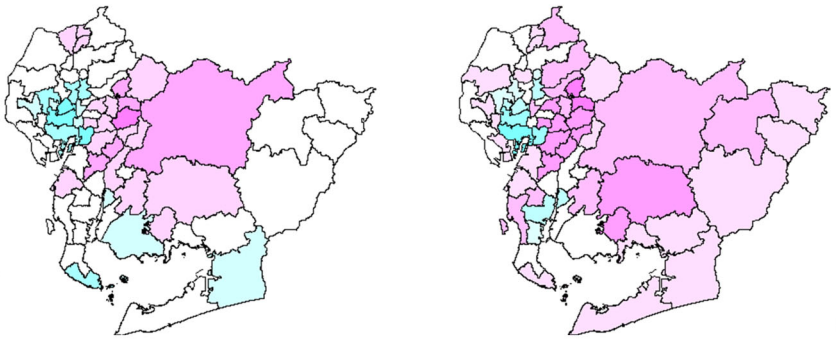
そのためには、まずデータを用意しなければいけません。何を使うかですが、ここでは例として、市区町村別の平均寿命をプロットしてみたいと思います。先ほどお話しした data.frame である sf オブジェクトに、このデータの列を追加していきます。

データの可視化②

- コロプレス図（Choropleth map）（続き）

```
LE.aichi <- read.csv("平均寿命_愛知県.csv")
map_jp.aichi <- cbind(map_jp.aichi,
                      LE.male=LE.aichi$平均寿命.男性,
                      LE.female=LE.aichi$平均寿命.女性)
plot(map_jp.aichi[, "LE.male"], pal=cm.colors)
plot(map_jp.aichi[, "LE.female"], pal=cm.colors)
```

sfオブジェクトにプロットしたいデータを追加

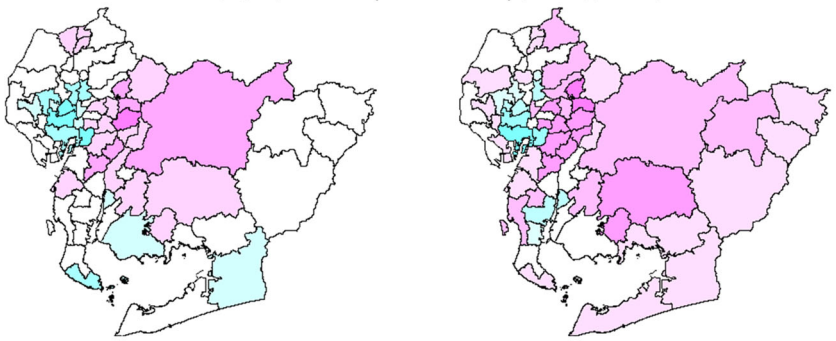


3

どのようにやるかというと、普通に data.frame を結合するので、cbind という関数を使って平均寿命.男性・平均寿命.女性の列を追加しています。これを追加してプロットすれば、このような図を非常に簡単に作ることができます。

データの可視化③

- コロプレス図（Choropleth map）（続き）



  - 平均寿命が相対的に長い地域（ピンク）・短い地域（水色）は、ある程度地理的に近接
  - 予測モデリングにおいても空間的な構造を考慮したい

19

これを見ていただくと、平均寿命が相対的に長い所はピンク色で、短い所は水色で示されていますが、ある程度地域的に近接していることが分かります。

このような空間的な相関をモデリングに当たっても反映していくことで、予測精度の高いモデルができるかと思いますので、うまくモデルの中に取り込んでいきたいということで、次に進みたいと思います。

隣接関係の把握

- 空間的な相関をモデル化するにあたり、地域間の空間的な関係に関する情報をモデルに与える必要がある
- spdepパッケージにより、隣接関係に関する情報が生成できる

```
library(spdep)
```

20

取り込むに当たっては、空間的な関係をモデルに与えてやる必要があります。実際にどのようにやるかというと、spdep パッケージというものがありまして、これを使ってやると、隣接関係に関する情報を作ることができます。

具体的には、次のページになります。

隣接関係の把握②

- spdepパッケージ
  - poly2nb()：各地域について、隣接する地域を要素として含むリストを作成

```
jpn.aichi.adjnum <- poly2nb(map_jp.aichi)
head(jpn.aichi.adjnum)
```

```
[[1]]
[1] 2 6 7 13 15 16
[[2]]
[1] 1 3 6 13
[[3]]
[1] 2 4 6 13 22 49 55
[[4]]
[1] 3 5 6 48 49
[[5]]
[1] 4 6 10 48 52 58
[[6]]
[1] 1 2 3 4 5 7 9 10
```

名古屋市千種区（1番目の地域）は、
名古屋市東区（2）、中区（6）、昭和区（7）、
守山区（13）、名東区（15）、天白区（16）と隣接

21

例えば、poly2nb という関数がありますが、これを使ってやると、sf オブジェクトの中に含まれるそれぞれの地域について、隣接する地域を要素として含むリストを作ることができます。その中身としてサンプルを出していますが、sf オブジェクトのなかでは名古屋市千種区が先頭にあり、リストの先頭にもこの千種区に関する情報が含まれています。千種区は、その下にある 2 番目、6 番目、7 番目、13 番目、15 番目、16 番目の地域と隣接しています。右側の図がその周辺の地図ですが、見ていただくと、確かにうまくいっていることが分かるかと思います。下の項目についても同じです。

隣接関係の把握③

- spdepパッケージ（続き）
  - nb2WB(), nb2INLA(), nb2mat()：poly2nb()の結果を空間データのモデリングで用いられる形式に変換

```
jpn.aichi.adj <- nb2WB(jpn.aichi.adjnum)
str(jpn.aichi.adj)
```

```
List of 3
 $ adj    : int [1:342] 2 6 7 13 15 16 1 3 6 13 ...
 $ weights: num [1:342] 1 1 1 1 1 1 1 1 1 1 ...
 $ num    : int [1:69] 6 4 7 5 6 8 5 4 6 7 ...
```

poly2nb()の結果に含まれる要素を結合

各地域が隣接する地域の数

```
nb2INLA("jpn.aichi.txt", jpn.aichi.adjnum)
```

出力先ファイル名

["jpn.aichi.txt"の内容]

```
69
1 6 2 6 7 13 15 16
2 4 1 3 6 13
3 7 2 4 6 13 22 49 55
4 5 3 5 6 48 49
5 6 4 6 10 48 52 58
6 8 1 2 3 4 5 7 9 10
```

地域の数

poly2nb()の結果

これを基にして他の関数を使ってやると、いろいろな空間モデリングで使うような形式でのデータを作ることができます。例えば、このあとでお話しする INLA というモデルを使う場合には、nb2INLA という関

数を使って、一つ前のページの情報を INLA で使う形式のデータに直してやります。

INLA では、地図情報をテキストファルとして読み込む必要がありますが、そのテキストファイルを作るための関数が用意されています。

次のページも同じです。

隣接関係の把握④

- spdepパッケージ（続き）
  - nb2WB(), nb2INLA(), nb2mat()：poly2nb()の結果を空間データのモデリングで用いられる形式に変換

```
jpn.aichi.mat <- nb2mat(jpn.aichi.adjnum, style="B")
jpn.aichi.mat
```

```
  [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
1    0    1    0    0    0    1    1    0    0     0     0     0
2    1    0    1    0    0    1    0    0    0     0     0     0
3    0    1    0    1    0    1    0    0    0     0     0     0
4    0    0    1    0    1    1    0    0    0     0     0     0
5    0    0    0    1    0    1    0    0    0     1     0     0
6    1    1    1    1    1    0    1    0    1     1     0     0
```

→ 名古屋市千種区（1番目の地域）は、2, 6, 7, ...番目の地域と隣接

23

これもあるモデリング手法で使う地図情報を与えるのですが、この関数を使うと行列形式で隣接関係を表していることになります。横に見ていくと、1 番目は名古屋市千種区でしたが、1 が入っているのが隣接している所で、2 番目、6 番目、7 番目等の地域と隣接しているという情報を表しています。ということで、この関数を使えばこのようなことが簡単にできます。

3．空間データのモデリング

24

これを使って生成した隣接関係に関する情報を使って、次に、空間データのモデリングを考えていきたいと思います。

発生件数のモデリング

- 右表のデータに基づき、発生件数を予測するモデルを考える

M1: $y_i \sim \text{Poisson}(\lambda_i)$

$\log(\lambda_i) = \log(E_i) + \beta_0 + v_i$

$v_i \sim N(0, \tau_v^{-1})$

- ベイズ的手法であれば、β_0, τ_vに事前分布を仮定

実績件数　期待件数

ID	Y	expL
1	31	22.11
2	129	141.20
3	10	8.95
4	183	165.63
5	13	14.04
6	20	19.65
…	…	…
46	174	202.57

25

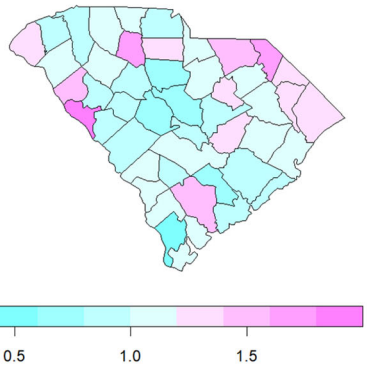
まず、発生件数のモデリングです。一旦、地図情報は忘れていただいて、このようなデータを持っていたとしましょう。全部で 46 件のデータがあって、それぞれ実績件数と期待件数が入っています。このようなデータがあったときに、実績件数をモデリングするとしたら、このような素直なモデルを考えると思います。オフセットとして期待件数を置いて、これが対数関数でポアソン分布のパラメータにつながっている。さらに、ベイズ的手法であれば、ここで使っているパラメータのβやτ_vに事前分布が仮定されている。そのようなモデルになります。

発生件数のモデリング②

- 何のデータ？
  - ある年のSouth Carolina州の各郡における呼吸器癌の件数

ID	region	name	Y	expL
1	1	abbeville	31	22.11
2	2	aiken	129	141.20
3	3	allendale	10	8.95
4	4	anderson	183	165.63
5	5	bamberg	13	14.04
6	6	barnwell	20	19.65
…	…	…	…	…
46	46	york	174	202.57

実績件数÷期待件数



- モデリングにあたり空間的な相関を考慮したい

26

実際にこのデータは何だったかというと、ある年の、アメリカのサウスカロライナ州の各郡における呼吸器がんの発生件数でした。これをプロットしてみるとこのようになります。実績件数と期待件数の比をプロットしているのですが、ここで言っている期待件数は、州全体での人口当たりの発生件数に各郡の人口を掛けたものです。これと実績件数の比を取った Standardized Incidence Ratio（SIR）をプロットしたものが、右側の図になります。若干例外はありますが、SIR の高いピンクの地域がおおむね固まっていることが見て取れると思いますので、これを考慮してモデリングをしていきたいと思います。

発生件数のモデリング③

- 例えば以下のようにして空間的相関を考慮
  - ICAR（Intrinsic CAR, CAR=Conditional AutoRegressive）

M2: $y_i \sim \text{Poisson}(\lambda_i)$

$$\log(\lambda_i) = \log(E_i) + \beta_0 + v_i + u_i$$

（u_i に ICAR）

$$u_i \sim N\left(\bar{u}_{\delta_i}, \tau_u^{-1}/n_{\delta_i}\right)$$

  - δ_i ： i番目の地域の近傍（`pol2nb()`で得られる隣接地域等）
  - $\bar{u}_{\delta_i}$：δ_iに含まれる地域におけるuの平均
  - n_{δ_i} ： δ_iに含まれる地域の数

27

先ほどセッション C-3 でも同じような話があったかと思いますが、一つのやり方としては、先ほどの M1 のモデルの中に、もう一つ追加の項として空間的な相関を表す項を追加してやる方法があります。このu_iは正規分布に従う確率変数で、正規分布の平均が隣接している地域のuの平均となっています。分散は何かというと、隣接している地域のuの分散の平均です。このuのようなモデルのことを ICAR モデルと呼んでいます。このように考えれば、地理的な関係をモデリングの中に簡単に取り込むことができます。

発生件数のモデリング④

- さらに年度別のデータが得られた場合

ID	region	name	year	Y	expL
1	1	abbeville	1	31	22.11
2	1	abbeville	2	26	22.06
3	1	abbeville	3	30	21.97
4	1	abbeville	4	27	21.94
5	1	abbeville	5	27	21.91
6	1	abbeville	6	33	21.86
…	…	…	…	…	…
276	46	york	6	209	227.18

$$\text{M3: } \log(\lambda_i) = \log(E_i) + \beta_0 + v_i + u_i + \beta_1 t$$

$$\text{M4: } \log(\lambda_i) = \log(E_i) + \beta_0 + v_i + u_i + \gamma_j^{IID} \sim N\left(\beta_1 t, \tau_\gamma^{-1}\right)$$

$$\text{M5: } \log(\lambda_i) = \log(E_i) + \beta_0 + v_i + u_i + \gamma_j^{RW} \sim N\left(\gamma_{j-1}^{RW}, \tau_\gamma^{-1}\right)$$

$$\text{M6: } \log(\lambda_i) = \log(E_i) + \beta_0 + v_i + u_i + \gamma_j^{RW} + \psi_{ij} \sim N\left(0, \tau_\psi^{-1}\right)$$

28

さらに、このような空間の情報だけでなく年度別の件数の情報もあったとしましょう。一つ目について 6 年分のデータがあって、最後の 46 個目の地域まで 6 年分あります。合計で 46×6 で 276 件のデータがあります。このようなときに、時間的な相関も考慮するようにモデルを拡張することも、実は簡単にできます。先ほどの M2 のモデルに対して、例えば M3 のモデルであれば、線形のトレンドを追加しています。M4 のモデルであれば、線形のトレンドに対して、更にランダムな要素も追加しています。M5 のモデルは、トレンドの代わりにランダムウォークを入れています。M6 であれば、ランダムウォークに加えて、更に空間的相関と時間的相関の交互作用項も入れています。このようなことで、モデル上は、簡単にこれを拡張していくことができます。

発生件数のモデリング⑤

- モデルM1～M6は、いずれも潜在ガウスモデルの一種

$$\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_1 \sim \prod_{i \in \mathcal{I}} \pi(y_i|x_i, \boldsymbol{\theta}_1)$$

$$\mathbf{x}|\boldsymbol{\theta}_2 \sim N\left(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \mathbf{Q}^{-1}(\boldsymbol{\theta}_2)\right)$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \sim \pi(\boldsymbol{\theta})$$

y_Iは条件付き独立かつx_{-i}に依存しない

モデルM1～M6では、$y_i|x_i, \boldsymbol{\theta}_1 \sim \text{Poisson}(\lambda_i)$

$$\log \lambda_i = \eta_i = \log E_i + \beta_0 + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)}$$

$\mathbf{x} = (\eta, \beta_0, \boldsymbol{\beta}_j, \boldsymbol{f}_k)$は$\boldsymbol{\theta}_2$の条件付きで多変量正規分布に従う
次元は標本数に比例して大きくなる

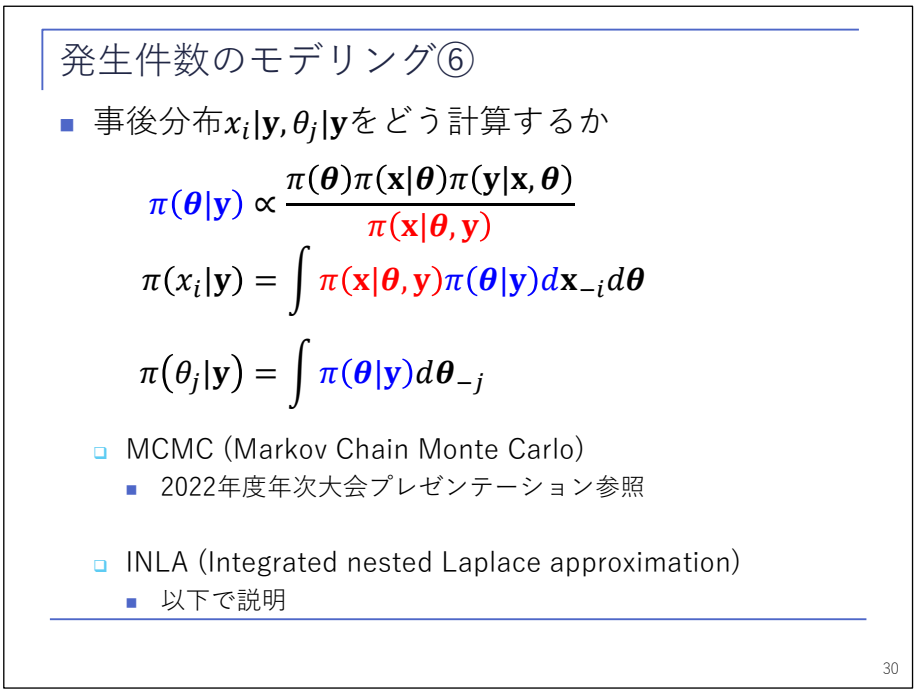
$\boldsymbol{\theta}=(\tau_v, \tau, \tau_\psi, \dots)$（ハイパーパラメータ）は正規分布ではない
一般に低次元（M1：1～M6：4）

29

M1 から M6 は、実は同じような形でこのような式で表すことができます。$\mathbf{y}$が観測値で、y_iが、x_i、$\boldsymbol{\theta}_1$の条件付きで互いに独立だという前提を置いています。y_iは、見ていただくとx_i以外の$\boldsymbol{x}$には依存していません。この$\mathbf{x}$は何かというと、2 行目の式のように表されまして$\boldsymbol{\theta}_2$の条件付きで多変量正規分布に従うような潜在変数です。$\boldsymbol{\theta}$はが$\boldsymbol{\theta}_1$、$\boldsymbol{\theta}_2$を並べたものですが、これがハイパーパラメータで、何らかの分布に従うのですが、こちらは正規分布である必要はありません。このようなモデルを潜在ガウスモデルといいます。

これまで考えてきたモデルで言うと、具体的にはy_iはポアソン分布に従います。ポアソン分布の期待値が二つ前のスライドにあるような式です。このように、オフセット項や切片、線形項、非線形の項、誤差項などの線形結合で表わされるηを何らかのリンク関数で、この場合は対数関数で変換したものがλになっています。

$\mathbf{x}$の次元は、データ量に応じて相当大きくなっていきます。10 万、20 万などの大きな次元になることがよくあります。それに対して$\boldsymbol{\theta}$は、一般に非常に次元が低くて、先ほど出てきた M1 のモデルであればハイパーパラメータは一つしかありませんし、一番複雑な M6 のモデルでも四つしかないということで、$\mathbf{x}$に比べれば非常に次元が低いという性質を持っています。



このモデルが決まれば、数式に従って、これと観測データを使って、このパラメータの事後分布を計算することができます。

式で表せばこのとおりなのですが、実際にこれを計算しようと思うと大変で、一般には解析解が得られないので、シミュレーションで計算するか、近似的に計算するかといった方法が取られます。シミュレーションで計算する方法として MCMC があるのですが、そちらについては前年度のプレゼンテーションで詳しくお話ししているので割愛いたします。もう一つ、近似的に計算する方法として INLA というものがありますが、今回はこちらについて細かくお話ししたいと思います。

INLA

- INLA (Integrated nested Laplace approximation)
  - 潜在ガウスモデルについて近似的な事後分布の計算を行う手法
  - ラプラス近似（モードの周りでの正規分布による近似）と数値積分を併用
  - MCMCに比べ高速で精度も高い
  - 詳細は以下を参照

    佐野誠一郎 (2022)「INLAによる時空間の従属性を考慮した頻度モデル」（2022年度優秀論文）

31

INLA とは何の略かというと、ここに書いてあるとおりなのですが、日本語にすれば「積分入れ子型ラプラス近似」という名前になるかと思います。先ほど出てきた潜在ガウスモデルについて近似的な事後分布の計算を行う手法で、名前のとおりラプラス近似と数値積分を併用している手法です。

発生件数のモデリング⑤

- モデルM1～M6は、いずれも潜在ガウスモデルの一種

$$\mathbf{y}|\mathbf{x},\boldsymbol{\theta}_1 \sim \prod_{i\in\mathcal{I}} \pi(y_i|x_i,\boldsymbol{\theta}_1)$$

$$\mathbf{x}|\boldsymbol{\theta}_2 \sim N\left(\boldsymbol{\mu}(\boldsymbol{\theta}_2), \mathbf{Q}^{-1}(\boldsymbol{\theta}_2)\right)$$

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \sim \pi(\boldsymbol{\theta})$$

y_Iは条件付き独立かつx_{-i}に依存しない

モデルM1～M6では、$y_i|x_i,\boldsymbol{\theta}_1 \sim \text{Poisson}(\lambda_i)$

$\log\lambda_i = \eta_i = \log E_i + \beta_0 + \sum_j \beta_j z_{ij} + \sum_k f_{k,j_k(i)}$

$\mathbf{x} = (\eta, \beta_0, \boldsymbol{\beta}_j, \boldsymbol{f}_k)$は$\boldsymbol{\theta}_2$の条件付きで多変量正規分布に従う
次元は標本数に比例して大きくなる

$\boldsymbol{\theta}=(\tau_v, \tau, \tau_\psi, \ldots)$（ハイパーパラメータ）は正規分布ではない
一般に低次元（M1：1～M6：4）

29

先ほど出てきた潜在ガウスモデルはこのようなものでしたが、一般には$\boldsymbol{\theta}$は非常に次元が低く、また、潜在変数の分散の逆行列（$\mathbf{Q}$の逆行列）は、一般的にゼロが非常に多い疎行列になっています。このような特徴を持っていることが多いので、これを利用して非常に効率的な計算を行います。

INLA

- INLA (Integrated nested Laplace approximation)
  - 潜在ガウスモデルについて近似的な事後分布の計算を行う手法
  - ラプラス近似（モードの周りでの正規分布による近似）と数値積分を併用
  - MCMCに比べ高速で精度も高い
  - 詳細は以下を参照

    佐野誠一郎 (2022)「INLAによる時空間の従属性を考慮した頻度モデル」（2022年度優秀論文）

31

精度と計算速度を両立させるために、ラプラス近似の計算の中で更にラプラス近似を利用するということで、このような名前がついています。

高速と言っていますが、どのくらい速いかというと、例えば M1 のモデルであれば、MCMC だと 3 分ぐらいかかるのですが、INLA を使えば 3 秒ぐらいで終わってしまいます。データ量が増えてくれば、更にその差も広がってくるので、非常に高速です。詳細は、今日ここでお話しすると時間もありませんので、前年度、優秀論文を取った佐野さんの論文がありますので、こちらをぜひ参照いただければと思います。

今回は、実際にどのように計算するかに焦点を当ててお話ししたいと思います。

R-INLA

- Rでの計算には、パッケージINLAが利用できる
  - ウェブサイト：https://www.r-inla.org/home
  - インストール

```
install.packages("INLA",repos=c(getOption("repos"),
  INLA="https://inla.r-inla-download.org/R/stable"),
  dep=TRUE)
```

  - チュートリアル
    https://www.precision-analytics.ca/articles/a-gentle-inla-tutorial/
  - テキスト
    - Bayesian inference with INLA
      https://becarioprecario.bitbucket.io/inla-gitbook/index.html
    - Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny
      https://paula-moraga.github.io/book-geospatial/

32

実は R では非常に簡単に計算できまして、計算のためのパッケージ INLA というものが用意されていま

す。残念ながら CRAN には登録されていないのですが、r-inla のサイトからダウンロードして、インストールすることができます。使い方はチュートリアルも用意されていますし、もう少し詳しく知りたい方は、ウェブ上でテキストも公開されていますので、このようなものをご覧いただければ学ぶことができるかと思います。

R-INLA②

- SC州の各郡における呼吸器癌の件数のモデル
  - inla()による計算

```
result.SCresp.inla <- inla(
  <FORMULA>
  family = "poisson", data = SCresp, E = expL,
  control.compute=list(dic=TRUE, cpo=TRUE, waic=TRUE)
  )
```

  - <FORMULA>
    - 1+[説明変数名]+…+f(…)+…
      1：切片
      [説明変数名]： 固定効果
      f(…)：変量効果、空間相関、時系列相関等
    - 具体的には次頁のとおり
  - 事前分布
    - デフォルトでは、切片 $N(0,0^{-1})$, 固定効果の係数 $N(0,0.001^{-1})$, 精度 Gamma(1,0.00005)
      （次頁では `param=c(2,0.5)` としてGamma(2,0.5)に変更）

33

使い方は非常に簡単で、このような 5 行のコードで実行することができます。赤字になっている FORMULA というのが肝になっていまして、次のページにそれぞれ具体的に記載があるのですが、ざっくりとした説明では GLM と同じような書式です。説明変数を並べているような形式で、特徴としては、非線形の項や誤差項を f という関数を使って指定してやるといったところが、glm 関数とは違っています。また、上を見ていただくと、事前分布に関する情報が書いていないのですが、特に指定しない場合には、デフォルトの事前分布が使われます。デフォルトの事前分布は、INLA のドキュメントを見ていただければ確認できます。

各モデルの<FORMULA> ※region2はregionと同じ内容

M1	$\beta_0 + v_i$	Y~ 1+f(region, model="iid", param=c(2,0.5))
M2	$\beta_0 + v_i + u_i$	Y~ 1+f(region, model="iid", param=c(2,0.5)) +f(region2, model="besag", param=c(2,0.5), graph="SCcounty_graph.txt")
M3	$\beta_0 + v_i + u_i + \beta_1 t$	Y~ 1+f(region, model="iid", param=c(2,0.5)) +f(region2, model="besag", param=c(2,0.5), graph="SCcounty_graph.txt")+year
M4	$\beta_0 + v_i + u_i + \gamma_j^{IID}$	Y~ 1+f(region, model="iid", param=c(2,0.5)) +f(region2, model="besag", param=c(2,0.5), graph="SCcounty_graph.txt") +f(year, model="iid", param=c(2,0.5))
M5	$\beta_0 + v_i + u_i + \gamma_j^{RW}$	Y~ 1+f(region, model="iid", param=c(2,0.5)) +f(region2, model="besag", param=c(2,0.5), graph="SCcounty_graph.txt") +f(year, model="rw1", param=c(2,0.5))
M6	$\beta_0 + v_i + u_i + \gamma_j^{RW} + \psi_{ij}$	Y~ 1+f(region, model="iid", param=c(2,0.5)) +f(region2, model="besag", param=c(2,0.5), graph="SCcounty_graph.txt") +f(year, model="rw1", param=c(2,0.5)) +f(ID, model="iid", param=c(2,0.5))

FORMULA のそれぞれの細かい設定ですが、一番シンプルな M1 のモデルであれば、変数 region の値ごと、すなわち地域ごとに異なる値を取る確率変数を f で指定しています。中身は何かというと、引数として model に iid を指定しており、これは独立同一分布に従う正規分布を表しています。さらに引数 param で、正規分布のハイパーパラメータである分散の逆数の事前分布について、デフォルトではガンマ分布でパラメータが 1 と 0.00005 ですが、パラメータを 2 と 0.5 に設定しています。

M2 のモデルであれば、地理的相関を考慮するために、f をもう一つ追加して、region2 に対して、先ほど作った地理的情報（隣接関係を表すテキストファイル）を、graph という引数で与えています。また、M5 であれば、ランダムウォークを追加するために、model に rw1 というモデルを追加しています。この model については、上の方の M2 では besag とありますし、下は rw1 とあるのですけれども、中身は INLA のドキュメントを見ていただけば、詳細な解説があります。

R-INLA③

- どんなモデルが使えるのか
  - y_iの確率分布（尤度）　（family=...で指定）

```
> names(inla.models()$likelihood)
 [1] "poisson"                          "xpoisson"
...
```

  - $f_{k,j_k(i)}$のモデル　（model=...で指定）

```
> names(inla.models()$latent)
 [1] "linear"    "iid"       "mec"       "meb"       "rgeneric"
 ...
```

  - リンク関数　（control.family=list(link=...)で指定））

```
> names(inla.models()$link)
 [1] "default"  "cloglog"  "ccloglog"  "loglog"  "identity"
 ...
```

  - 詳細はinla.doc("^besag$", "latent")等で確認
    第１引数は正規表現なので完全一致とするよう^, $で囲む

35

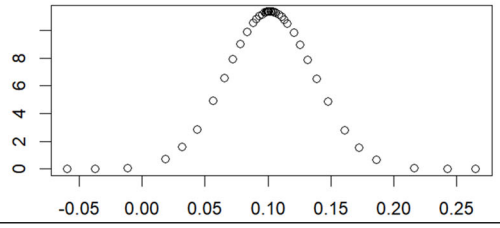
ドキュメントをどのように見るかですが、例えばどのようなモデルが指定できるかというところは、こちらにあるように、inla.models という関数があって、この結果を names 関数に入れてやると、リストが得られます。詳しく知りたい場合には、ここで分かった名前を inla.doc という関数の中に入れてやれば、更に細かい情報を得ることができます。この中で、例えば事前分布のデフォルトは何かという情報が得られます。

R-INLA④

- inla()の出力（inlaオブジェクト）に含まれる値

summary.fixed, summary.random, summary.hyperpar	要約統計量（切片・固定効果、変量効果、ハイパーパラメータ）
marginal.fixed, marginal.random, marginal.hyperpar	事後密度上の点(x, y)（切片・固定効果、変量効果、ハイパーパラメータ）
cpo$cpo	Conditional Predictive Ordinates (後述)
waic	WAIC, local WAIC, 実効パラメータ数
dic	DIC, local DIC, 実効パラメータ数

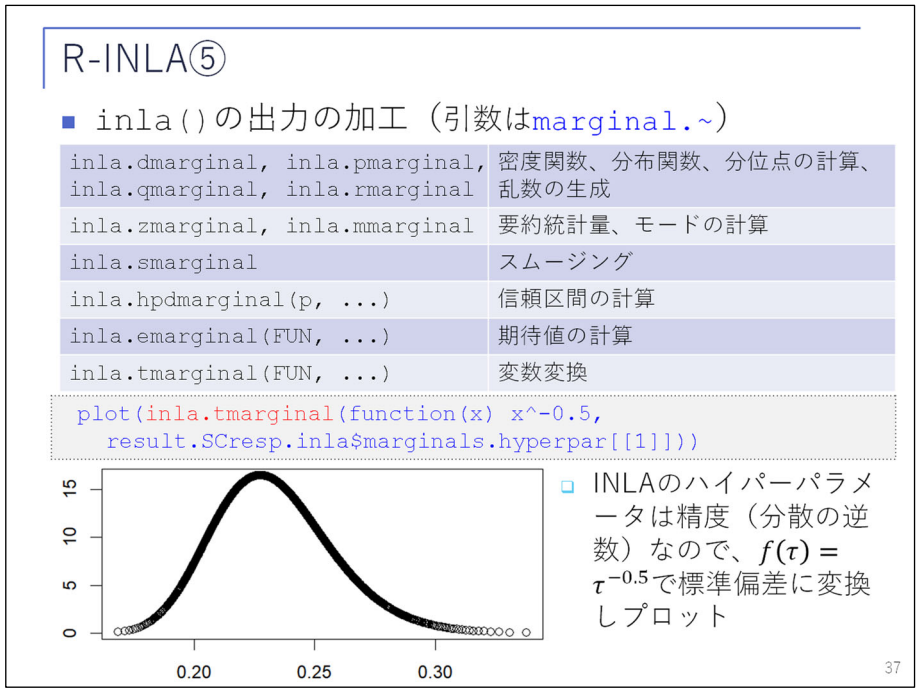
```
plot(result.SCresp.inla$marginals.fixed$`(Intercept)`)
```



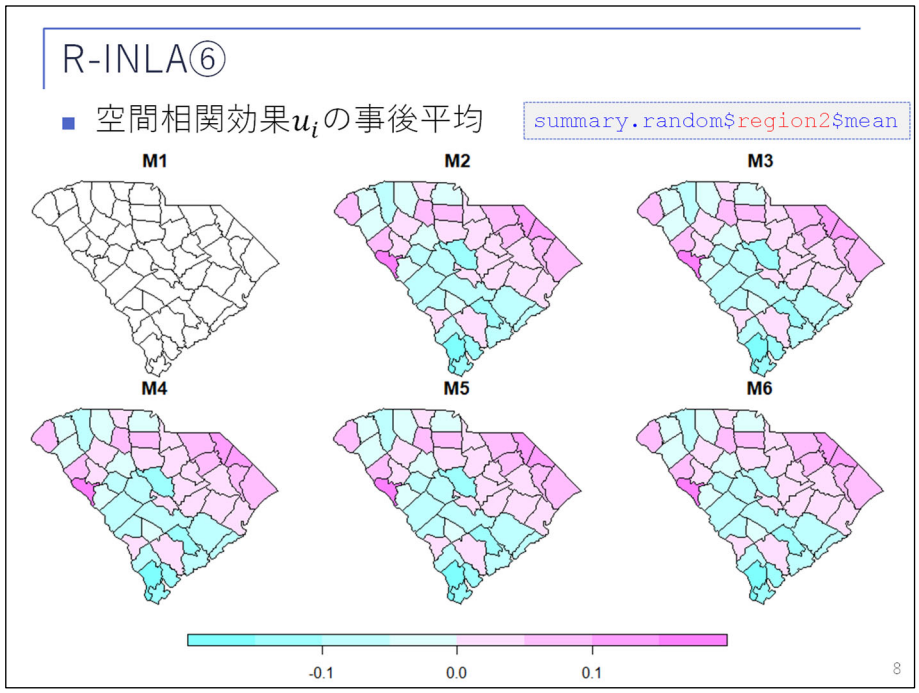
36

結果をどのように見るかですが、INLA パッケージの答えは、inla オブジェクトという形のデータ形式で返ってきます。この中にいろいろな変数が含まれていて、一番上の行に summary から始まる変数が幾つかありますが、パラメータの事後分布の結果の要約統計量。例えば期待値や、標準偏差、最頻値、中央値

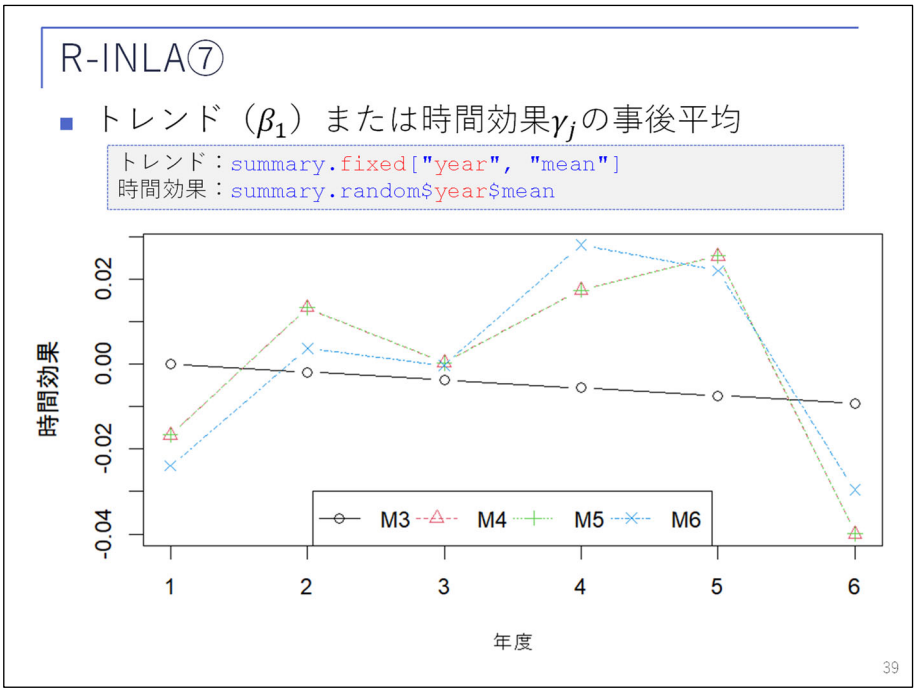
などの情報が入っています。二つ目の marginal で始まる変数には、それぞれの事後分布の密度関数の x と y=f(x)がプロットできるような情報が入っています。これをプロットしてやると、下のグラフのような密度関数のグラフが得られます。その下三つは、適合度の尺度なのですが、また後でお話しします。



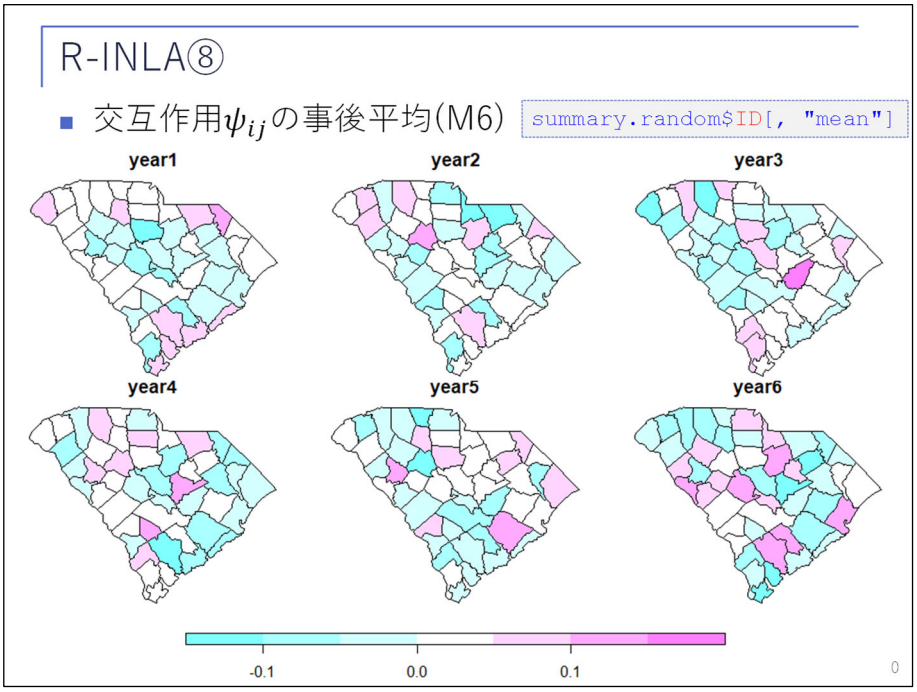
パッケージの中では、先ほどの marginal を加工していろいろな情報を得る関数もあるのですが、時間が足りないようなので、割愛させていただきます。



結果をプロットしたものが、こちらです。時期ごとの空間相関効果u_iを設定して、事後平均をプロットしています。それぞれのモデルごとに、先ほどお話しした region2 という変数に対する summary.random の効果の平均を表示しています。



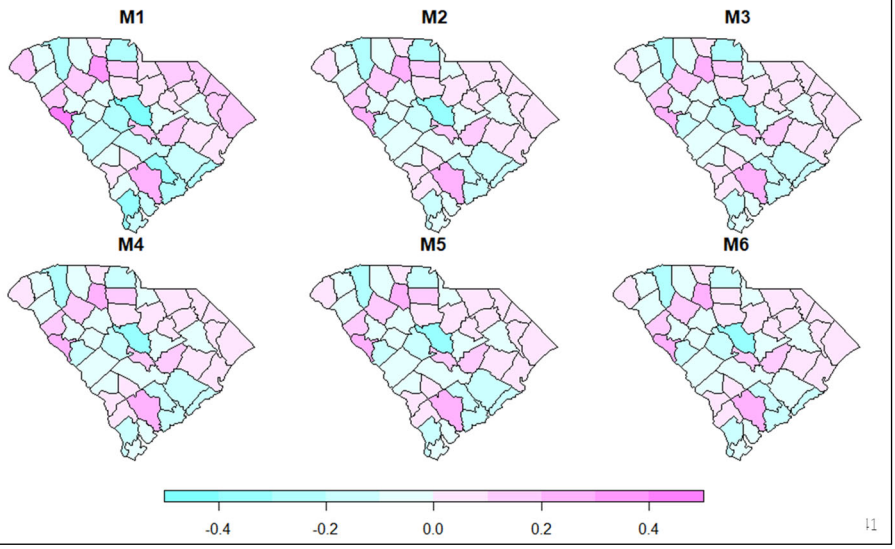
トレンドや時間効果についても同じように情報を得ることができます。トレンドについては、線形の項ですので、summary.fixed の中にあります。時間効果であれば、year という変数の random の効果の中にあります。スライドのグラフではこれの平均を表示しています。



このスライドでは、M6 のモデルに入っていた交互作用について事後平均を表示しています。これについては、後半の野村さんのパートでもプロットが出てくるかと思います。

R-INLA⑨

- 無相関の誤差項v_iの事後平均 `summary.random$region$mean`



41

このスライドは、最後に無相関の誤差項の事後平均のプロットです。これにおかしな偏りがないか確認することは、モデルの診断に使えるのではないかと思います。

R-INLA⑩

- 適合度の尺度
  - 観測ごと（地域・時点ごと）の適合度
    - CPO： 自身以外の観測に基づく事後密度　$\pi(y_i|\mathbf{y}_{-i})$
      大きいほうが適合度が高い
  - 全体的な適合度
    - WAIC, DIC
    - LMPL (Log Marginal Predictive Likelihood)：$\sum_i \log CPO_i$

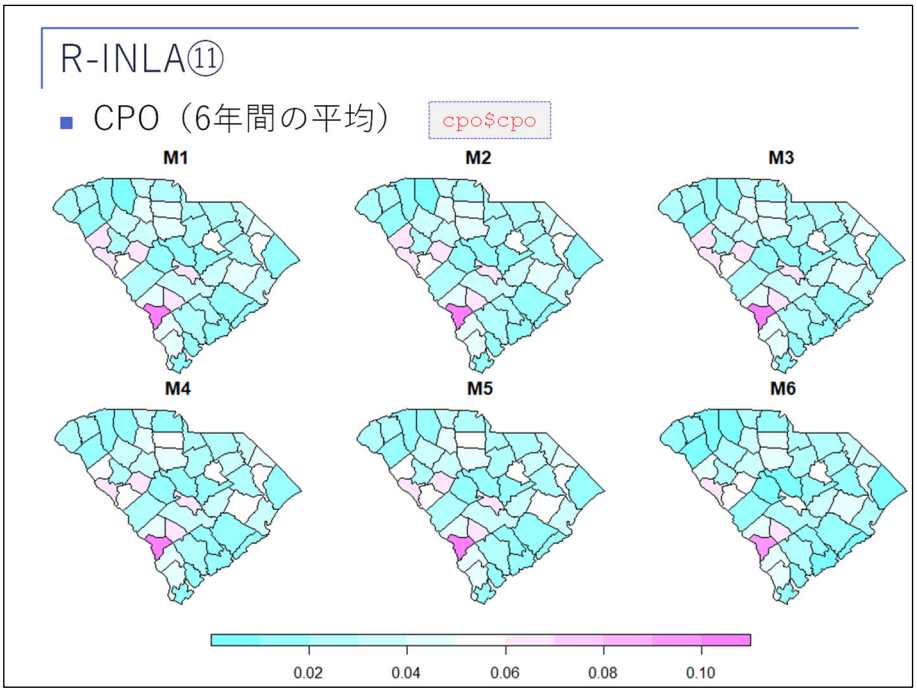
モデル	DIC	waic	LMPL
M1.UH	1998.33	2008.51	-1005.41
M2.UH+ICAR	1998.93	2009.21	-1005.81
M3.UH+ICAR+trend	2000.69	2011.49	-1007.06
M4.UH+ICAR+iid.trend	1996.06	2006.15	-1004.75
M5.UH+ICAR+rw	1996.04	2006.13	-1004.73
M6.UH+ICAR+rw+int	2030.39	1984.62	-1047.55

42

最後に適合度の尺度として、先ほど三つ変数があるという話をしましたが、INLA パッケージで利用できるものとして、観測ごとの適合度の尺度として CPO というものがあります。Conditional Predictive Ordinate の略ですが、自身以外の観測値を使って当てはめたモデルの事後密度であって、いわばクロスバリデーション的な手法といえるかと思います。これが大きいほど当てはまりがいいということを表しています。

一方で、全体的な適合度の尺度として、WAIC や DIC があります。これは、前年度のプレゼンテーション

でも出てきましたので省略します。もう一つ、LMPL があり、CPO を使って計算することができます。これもやはり同じように、大きい方が当てはまりがいいという尺度になります。これを比べた結果はスライドのとおりです。



最後に、CPO をモデルごとにプロットしたものがこちらの図です。inla オブジェクトの中に含まれる CPO の値をプロットしたものです。

最後は駆け足になってしまいましたが、以上が INLA パッケージの基本的な使い方です。

ここからは、実際の自賠責保険の統計データを使った分析例について、野村さんからご説明したいと思います。

【司会】 渡辺さん、ありがとうございます。引き続き野村さんからの発表を放映します。

【野村】 それでは、ここから野村に代わりまして、学生と実施した二つの解析例について紹介させていただきます。

4．空間時系列データのモデリング
～自賠責保険での解析例①～

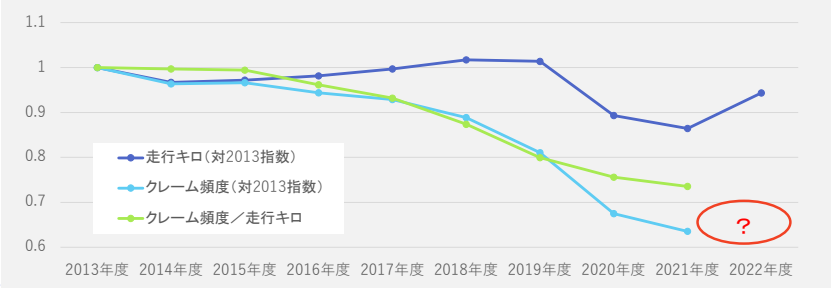
44

一つ目は、空間時系列データのモデリングを利用した、自賠責保険での解析例になります。

自賠責保険のクレーム頻度の推移

- 自賠責保険のクレーム頻度は近年減少トレンドにありコロナ禍の交通量減少に伴いさらに低下している
- 都道府県別支払件数の空間時系列データ解析により交通量回復後のクレーム頻度の変化を予測したい

自家用乗用自動車のクレーム頻度※1と走行キロ※2の推移



※1 損害保険料率算出機構「自賠責保険統計」よりクレーム頻度＝支払件数÷経過台数により算出
※2 国土交通省「交通関係統計資料 自動車燃料消費量調査」より集計

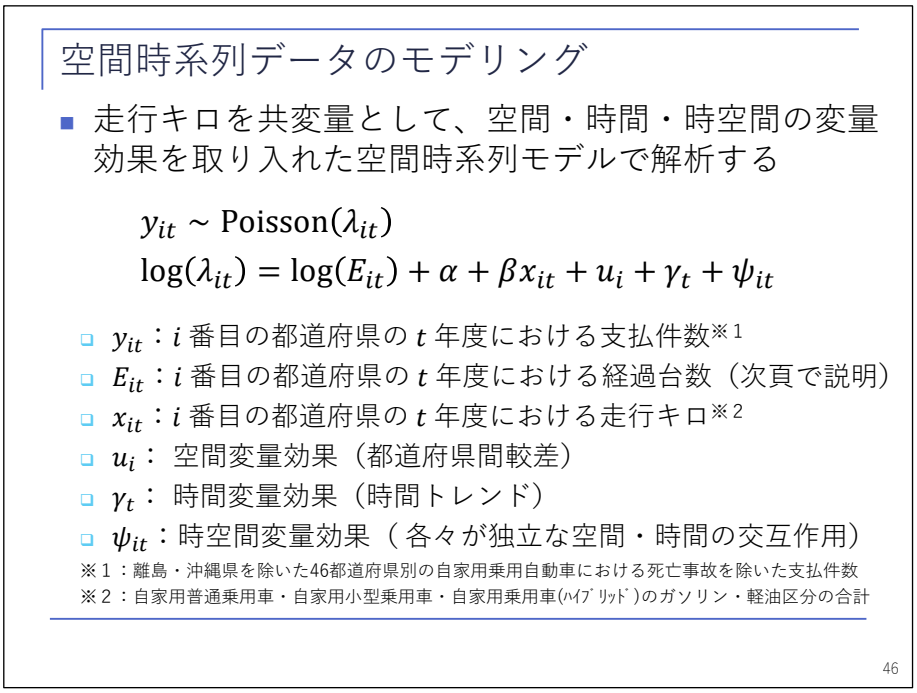
45

まずは、研究の目的から説明します。自賠責保険のクレーム頻度は近年減少トレンドにありまして、さらに、最近のコロナ禍に伴う交通量の減少に伴って、更に低下している状況です。下のグラフは、2013 年度から 2021 年度までの自賠責保険自家用乗用自動車におけるクレーム頻度の推移になります。水色が 2013 年対比の指数ですけれども、最近ぐっと下がっていることが分かります。

このぐっと下がっている所には、青線の国土交通省の自動車燃料消費量調整より集計した走行キロの推移が関係しておりまして、コロナ禍に入って大きく下がったことに伴って、クレーム頻度も低下しています。

クレーム頻度と走行キロの比を取りますと、このように緩やかな減少トレンドになることが分かります。

走行キロについては 2022 年度の数値が判明している状態で、クレーム頻度が 2022 年度に幾らになるかを予測しようということが、本研究の目的になります。特に、この走行キロが 2022 年度に回復してきていますので、その回復に伴ってクレーム頻度がどのように変化するかを予測しよう、ということになります。



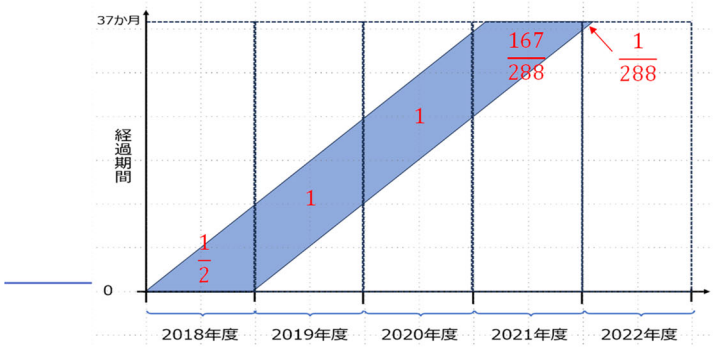
では、モデリングの内容に移ります。ただいまの走行キロを共変量としまして、空間・時間・時空間の変量効果を取り入れた空間時系列モデルを用いて解析します。まず、目的変数 y_{it} が、i 番目の都道府県における t 年度の支払件数となります。こちらは、離島、沖縄県を除いた 46 都道府県別の、自家用乗用自動車における死亡事故を除いた支払件数となっています。こちらを平均 λ_{it} のポアソン分布に従うと仮定したうえで、対数リンクを使って、$\log \lambda_{it}$ がこのような式でモデル化されています。最初の項はオフセット項で、i 番目の都道府県の t 年度における経過台数を表します。つまり、経過台数に比例して支払件数が増えるという構造を表しています。

経過台数の求め方は、次のページで説明します。

経過台数の推計

- 前頁の経過台数E_{it}（単位：台×年）は保険期間別の新契約台数を用いて推計した
- 新契約の保険始期が年度内で一様に分布する仮定の下で、下図の面積により各年度の経過台数が求まる

2018年度の新契約（保険期間37か月）1台あたりの年度別経過台数



47

経過台数については、料率機構の資料から保険期間別の新契約台数が与えられておりまして、そちらを用いて推計しました。新契約の保険始期が年度内で一様に分布する仮定の下で、経過台数が平均的に何台になるかを求めています。例えば 2018 年度における新契約、保険期間 37 か月のものについて、1 台当たりの年度別の経過台数がどのように推計されるか考えてみます。2018 年度の初めに保険期間が始まった場合、2018 年度は丸一年間保険期間に含まれることになるので、1 台×年という単位でカウントします。一方で、2018 年度の終わりから保険期間が始まると、2018 年度中は保険期間に含まれないので、0 台×年となります。それが 2018 年度中で一様に分布すると仮定しますと、間を取って 2 分の 1 となります。

このような考え方ですけれども、こちらの下の図に示すような、2018 年度を底辺として経過期間を高さとして斜めに進ませた平行四辺形の面積を年度別に取ることによってカウントすることができます。このような関係を利用して、保険期間別に 1 台当たり各年度の経過期間が幾らになるかということを集計することで、経過台数を推定しております。

空間時系列データのモデリング

- 走行キロを共変量として、空間・時間・時空間の変量効果を取り入れた空間時系列モデルで解析する

$$y_{it} \sim \text{Poisson}(\lambda_{it})$$
$$\log(\lambda_{it}) = \log(E_{it}) + \alpha + \beta x_{it} + u_i + \gamma_t + \psi_{it}$$

  - y_{it}：i 番目の都道府県の t 年度における支払件数※1
  - E_{it}：i 番目の都道府県の t 年度における経過台数（次頁で説明）
  - x_{it}：i 番目の都道府県の t 年度における走行キロ※2
  - u_i： 空間変量効果（都道府県間較差）
  - γ_t： 時間変量効果（時間トレンド）
  - ψ_{it}：時空間変量効果（各々が独立な空間・時間の交互作用）

※1：離島・沖縄県を除いた46都道府県別の自家用乗用自動車における死亡事故を除いた支払件数
※2：自家用普通乗用車・自家用小型乗用車・自家用乗用車(ﾊｲﾌﾞﾘｯﾄﾞ)のガソリン・軽油区分の合計

46

以上が経過台数で、その次が切片、その次が回帰係数 β ×共変量の走行キロとなります。こちらは、自家用乗用自動車に対応すると思われる、ガソリン・軽油区分の合計値を示しています。あとは変量効果でして、空間の変量効果 u_i は、都道府県間の較差を表します。次が、時間の変量効果 γ_t で、こちらは時間のトレンドを表す項になります。最後に ϕ_{it} は、時空間の変量効果になりまして、おのおのが独立と仮定している空間・時間の交互作用を表す項となります。

モデル選択

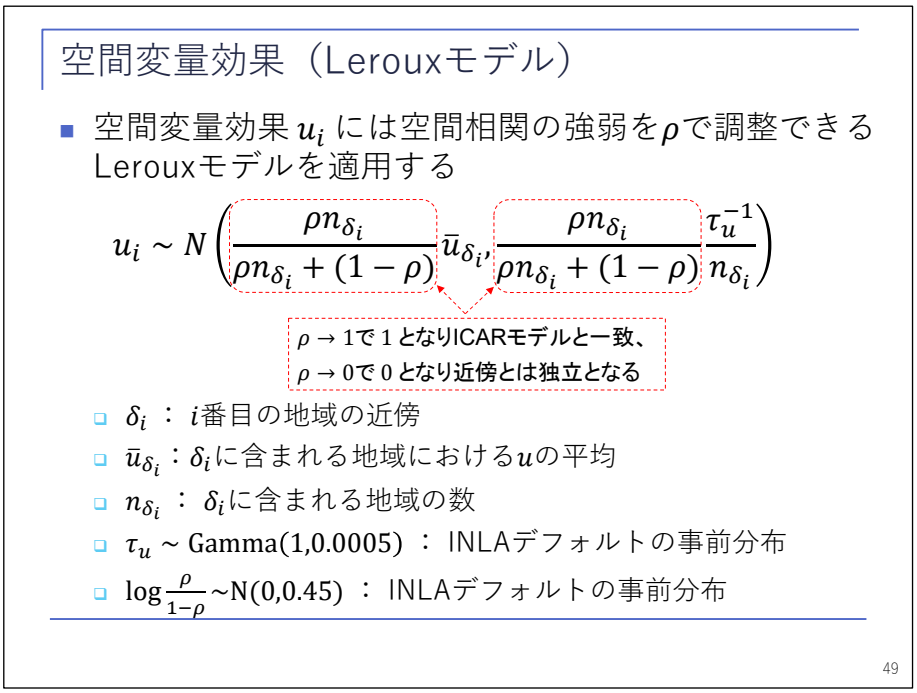
- 全体的な適合度の尺度（DIC, WAIC, LMPL）を比較した結果、ここまで紹介した全ての変量効果を取り入れた採用モデルが最良となった
- 走行キロに対する回帰係数は推定値 0.695（95%ベイズ信頼区間 0.213～1.107）となり、クレーム頻度に有意な効果をもたらしている

モデル	DIC	WAIC	LMPL
採用モデル $\log(\lambda_{it}) = \log(E_{it}) + \alpha + \beta x_{it} + u_i + \gamma_t + \psi_{it}$	5265.1	5138.0	-2892.3
採用モデル－時空間変量効果ψ_{it} $\log(\lambda_{it}) = \log(E_{it}) + \alpha + \beta x_{it} + u_i + \gamma_t$	6418.0	7414.0	-3362.7
採用モデル－時間変量効果γ_t＋線形トレンド$\beta_t t$ $\log(\lambda_{it}) = \log(E_{it}) + \alpha + \beta x_{it} + u_i + \beta_t t + \psi_{it}$	5273.4	5149.9	-2924.1

48

まず前ページのモデルと他の候補として考えたモデルをモデル選択の尺度により比較いたします。前ページのモデルと、そこから時空間変量効果を抜いたモデル、および、時間変量効果を抜いて代わりに線形のトレンド項を設けたモデル、この三つで比べてみますと、DIC、WAIC、LMPL ともに前ページのモデルが

一番よいという結果になっているため、前ページのモデルを採用モデルといたします。また、このモデルで走行キロに対する回帰係数を求めますと、推定値は 0.695 となります。さらに、95%の信頼区間を求めますと、0.213～1.107 と有意に正となるわけですね。ですので、走行キロはクレーム頻度に有意な効果をもたらしているといえます。



ここから各変量効果の説明に入ります。空間変量効果については、ICAR モデルの発展版である Leroux モデルを用いています。こちらは、ICAR モデルに対して、空間相関の強弱を調整できる ρ というパラメータをこのような形で組み込んでいるモデルとなっています。 ρ を 1 に近づけると、赤枠で囲っている部分を除けば ICAR モデルそのものとなっていますので、ICAR モデルと一致します。一方で、 ρ を 0 にもっていくと、こちらは 0 に近づいていきまして、周辺との相関がなくなって最終的には各地域独立のモデルになります。 ρ を 0 と 1 の間で調整することで、空間相関の強弱を調整できるモデルとなっています。

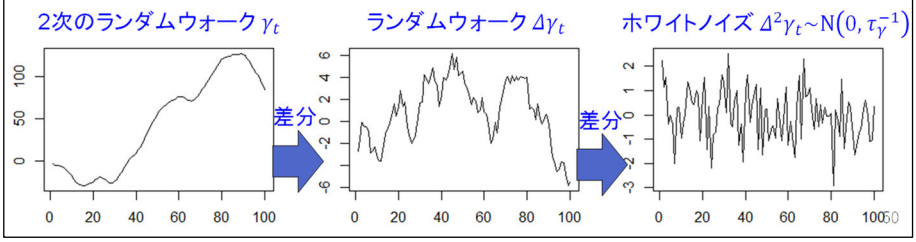
分散を調整するパラメータ τ_u や ρ については下に示す事前分布を仮定しています。いずれも INLA デフォルトで設定されている事前分布となっています。

時間変量効果（2次のランダムウォーク）

- 時間変量効果 γ_t には増減トレンドを保ちながら推移する2次のランダムウォークを適用する

$$\Delta^2\gamma_t = \Delta\gamma_t - \Delta\gamma_{t-1} = (\gamma_t - \gamma_{t-1}) - (\gamma_{t-1} - \gamma_{t-2}) \sim \mathrm{N}\left(0, \tau_\gamma^{-1}\right)$$

  - Δ ： 階差オペレータ（前時点からの変化量をとる）
  - $\tau_\gamma \sim \mathrm{Gamma}(1, 0.0005)$：INLAデフォルトの事前分布



続いて、時間の変量効果は、2 次のランダムウォークと呼ばれる、　このように滑らかなトレンドを持って推移するモデルを仮定しています。

なぜ滑らかに推移するかといいますと、前の時点との差分を取りますとこのような普通のランダムウォークになるためです。

ランダムウォークは、前時点からの差分を取るとホワイトノイズ、すなわち各時点で完全に独立のモデルに、同じ分布に従うモデルとなります。

ですので、ランダムウォークは、少し上がったら、しばらくガタガタ動いて、前の時点の値をある程度保存します。それを各時点の差分にするということは、ランダムウォークに従って、前時点の変化量が少しずつ変わっていきます。前時点の傾きが少しずつ変わっていくというトレンドをランダムウォークという形で保持しながら推移するモデルとなります。時間変量効果はこちらのモデルを仮定してトレンドをもって推定を予測する方針を取ります。

時空間変量効果（独立同一分布）

- 時空間変量効果 ψ_{it} は i,t の組合せ毎に独立に正規分布に従うと仮定する

$$\psi_{it} \sim \mathrm{N}\left(0, \tau_{\psi}^{-1}\right)$$

$$\tau_{\psi} \sim \mathrm{Gamma}(1, 0.0005)$$

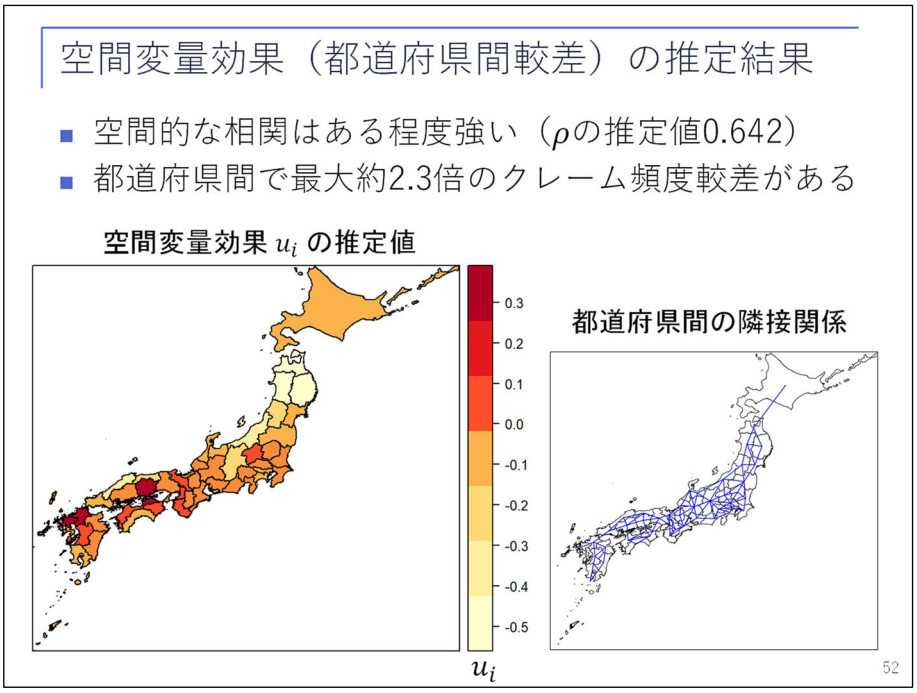
- 時空間変量効果は時間変量効果＋空間変量効果で表現できない時間・空間の間の交互作用項である（推定値に何らかの系統的特徴が見られるなら独立性の仮定が崩れるためモデル改良の余地ありとなる）
- データ毎に独立にポアソン分布の平均 λ_{it} を撹乱させることから、データの分散を実質的にポアソン分布より大きくさせる効果をもつ（過分散モデル）

51

最後に時空間の変量効果は、ｉ と ｔ の組合せ毎に、独立に正規分布に従うモデルを仮定します。意味合いとしては、時間の変量効果と空間の変量効果で表現できない時間と空間の間の交互作用項になります。

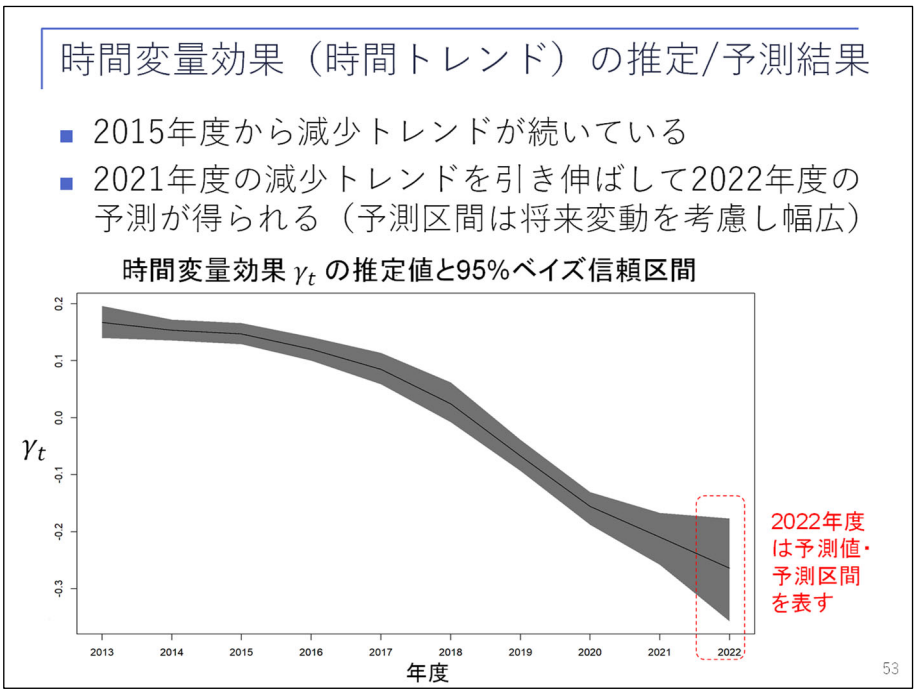
おのおの独立と仮定していますけれども、推定した結果、その推定値自体に何らかの系統的な特徴が見られることがあります。そのような場合は、独立性という仮定が崩れてしまいますので、何かしらモデル改良の余地があるのではないかと考えることができます。

また、独立な変量効果を入れる意味合いとしては、データ毎に独立に値を取りますので、ある意味ポアソン分布の平均 λ_{it} をランダムにかく乱させる効果をもたらします。ポアソン分布の分散は本来は非常に小さいのですが、その分散に変量効果によるばらつきが加わることで、実質的にポアソン分布よりも大きい分散を持つデータをモデル化できるという効果があります。このようなモデルを「過分散モデル」といいます。



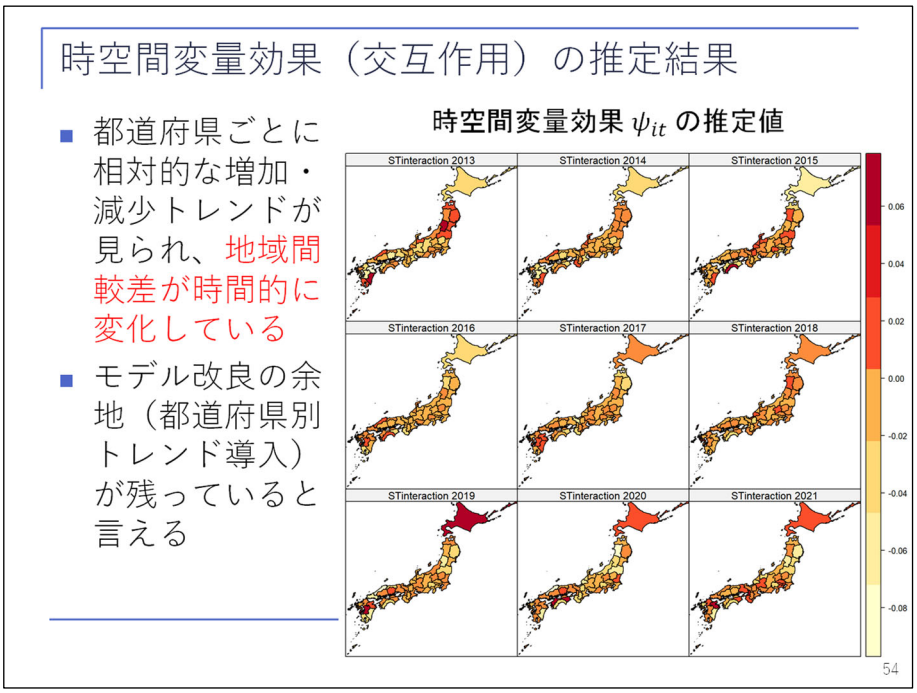
以上がモデルの説明になりまして、ここからは推定結果の説明に移ります。

まずは、空間変量効果の推定結果です。推定に当たっては、右下にあるような都道府県間の隣接関係を用いております。主に陸地が接していたり、道路がつながっている場合に隣接させていて、その隣接する地域間には割と強く相関があるという結果になっています。左側の図が、空間変量効果の推定値をカラースケールで表したものです。低め、やや高め、高めといった地域がところどころ出てきていて、空間的な相関の強さを表す ρ の推定値は 0.642 となっております。また、都道府県間の較差は大きく、低い都道府県と高い都道府県で比べますと、最大約 2.3 倍ものクレーム頻度の較差がございます。



続きまして、時間の変量効果です。

推定値をこちらの図に示しています。灰色の帯は、95%ベイズ信頼区間になります。2015 年あたりから顕著な減少トレンドが始まっていて現在まで続いています。2021 年から 2022 年度へとトレンドを延ばして予測を行っています。2 次のランダムウォークは、トレンドをそのままランダムウォークという形である程度保持できるものですから、予測する際にもそのトレンドを保持するのですが、ランダムウォークのばらつきを考慮してベイズ信頼区間の帯が広がる形になります。2022 年度については予測値と予測区間を表しています。このように 21 年から 22 年にかけても、その前の年のトレンドを引き継いで減少するという形で予測されています。



最後に、時空間の変量効果（交互作用）の推定結果です。

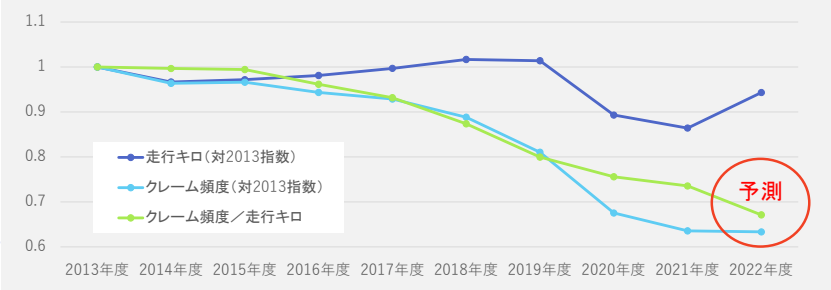
左上から右下にかけて各年の推定値を色で示していますが、都道府県ごとに相対的に増加・減少しているようなトレンドが見られます。顕著なものですと、北海道は薄い色からだんだん濃い色になっていて、他の都道府県に比べて相対的に、若干ではあるのですけれども、相対的に増えています。

逆に相対的に減っているような県などもありまして、全ての都道府県でクレーム頻度は減少トレンドにはあるのですが、その減少の速さに都道府県間で較差があることが分かりました。ですから、このモデルにもまだ改良の余地があって、例えば都道府県別のトレンドを導入することで、より当てはまりがよくなるかもしれないといった課題が残っているといえます。

自賠責保険のクレーム頻度の予測結果

- 2022年度のクレーム頻度を予測すると、走行キロの回復効果と減少トレンドの相殺でほぼ横ばいとなった
- この予測後に2022年度の自賠責保険統計が公表され、2022年度の自家用乗用自動車の支払件数実績は前年度からほぼ横ばい（152件減の375,316件）であった

自家用乗用自動車のクレーム頻度と走行キロの推移と2022年度予測



55

以上のモデルを使って 2022 年度の全国のクレーム頻度を予測しました。

最初にお見せしたグラフを 2022 年度まで引き延ばして予測しまして、結果としては、先ほど見ていただいた時間変量効果の減少トレンドと走行キロの回復によるクレーム頻度の上昇がほぼ相殺された形になり、2021 年度から 2022 年度にかけて水色のクレーム頻度はほぼ横ばいか、微減するような結果になりました。

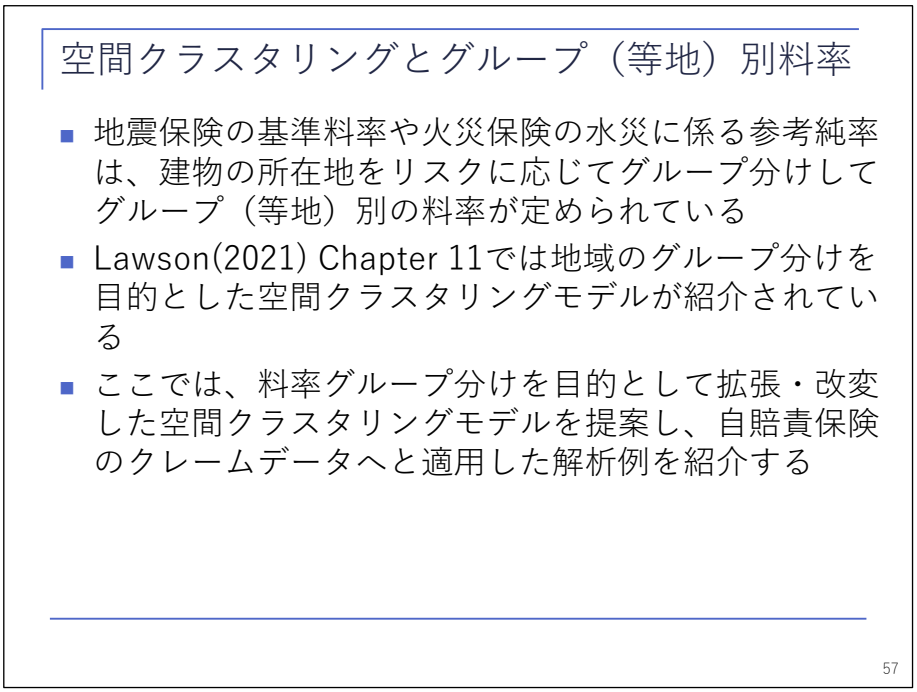
こちらの分析を行ったあと、9 月の末頃に 2022 年度の自賠責保険統計が料率機構より公表されました。

その結果、クレーム頻度ではまだ見られていないのですが、支払件数としては 2021 年度から 22 年度にかけて 152 件減の微減という、私たちの予測にほぼ近い結果になりました。

5．空間データクラスタリング ～自賠責保険での解析例②～

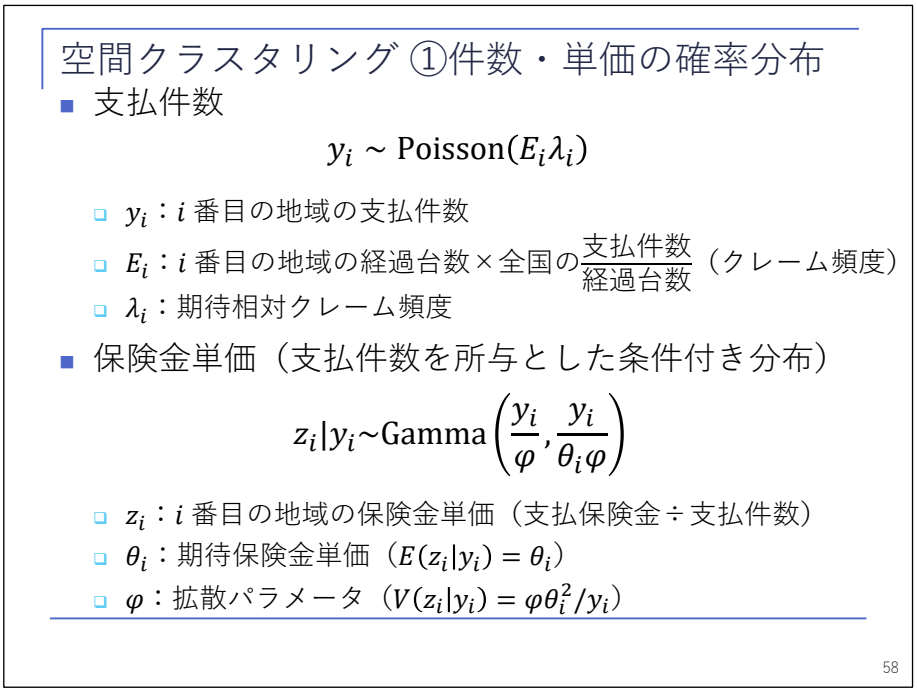
56

二つ目の解析例は、こちらも自賠責保険を扱い、空間データクラスタリングという手法を使った例を紹介します。



まず、背景として、現行でも地震保険の基準料率や火災保険の水災に係る参考純率などでは、建物の所在地ごとにリスクに応じたグループ（ランク）が決まっておりまして、そのランク（等地）別の保険料率が定められております。このように、都道府県やもっと細かい地域ごとに保険料を分けることは実務上でも行われておりまして、さらに、今回紹介しております本の Chapter 11 でも、地域のグループ分けを目的とした空間クラスタリングのモデルが紹介されています。

ここでは、地域別料率のグループ分けを目的として、Lawson の本のモデルを拡張・改変した空間クラスタリングのモデルを提案し、そのモデルを自賠責保険のクレームデータへと適用した解析例を紹介したいと思います。



ここでは、支払件数に加えて保険金単価もモデル化します。

支払件数 y_i は、このようなポアソン分布に従います。

E_i は、i 番目の地域における経過台数に掛けることの、全国のクレーム頻度、すなわち「経過台数」分の「支払件数」です。すなわち、クレーム頻度が全国一定のものであった場合に、i 番目の地域でどのくらいのクレームが発生するかを表しています。そこに λ_i を掛けていて、こちらが地域間の相対的なクレーム頻度の較差を表す潜在パラメータになります。

さらに保険金単価は、支払件数 y_i を与えられたもとで、このようなガンマ分布に従うという形でモデル化しております。この中に出てくる θ_i が、期待保険金単価となります。さらに、分母に出てくる ϕ は、分散を調整するためのいわゆる拡散パラメータといわれるものになります。

このようなパラメータを持った分布をもって、まず支払件数と保険金単価をモデル化します。

そのうえで、期待相対クレーム頻度 λ_i および期待保険金単価 θ_i について、次のような形でモデル化します。

空間クラスタリング ②期待頻度・期待単価モデル

- 前頁の期待相対クレーム頻度・期待保険金単価それぞれを以下のモデルでグループ分けして推定する

$$\log \lambda_i = \alpha_1 + \sum_{g=1}^{G_1} \left(g \times \beta \times w_{1,i,g} \right)$$

$$\log \theta_i = \alpha_2 + \sum_{g=1}^{G_2} \left(g \times \beta \times w_{2,i,g} \right)$$

頻度・単価ともグループgが1上がる度に期待値がexp(β)倍される（共通の較差）

- G_1, G_2：期待相対クレーム頻度/期待保険金単価のグループ数
- β：期待相対クレーム頻度/期待保険金単価のグループ間較差
- $w_{1,i,g}$：期待相対クレーム頻度のグループgへの所属割合(重み)
- $w_{2,i,g}$：期待保険金単価のグループgへの所属割合(重み)

59

まず、期待相対クレーム頻度 λ_i と期待保険金単価 θ_i それぞれに対数リンクを取って、右辺にはそれぞれの切片項＋＋このような項を設けます。

こちらがグループ化のキーとなる部分になります。この β は、二つのモデルの間で共通のパラメータとして用意しております。そこに、それぞれのグループに所属する割合となる重みを掛けていて、この g がグループ番号です。グループ番号×β×重みといった形で、これを全てのグループで合計しています。

どのような意味かといいますと、グループが一つ上がるごとに、掛かってくるものが β だけ増えるのです。グループ 1 に 100%所属している場合は「1×β×100%」で、全部で β になります。グループ 2 に 100%所属している場合は、g が 2 のときに 100%になりまして、「2×β×100%」で 2β となります。ということで、所属するグループが一つ上がるたびに、$\log\lambda$ あるいは $\log\theta$ が β 増えて、その期待値が、exp(β)倍されることになります。そのような意味で、β は重要な役割を果たしていて、クレーム頻度と保険金単価の共通のグループ間較差を表すものになっています。ただし、お互いにグループ数を同じにする必要はなく、G_1、G_2 のように、グループ数はデータに応じて調整して、異なるグループ数にしています。

空間クラスタリング ③グループ所属割合のモデル

- 期待相対クレーム頻度($k=1$)・期待保険金単価($k=2$)のグループ所属割合をそれぞれ以下のモデルで定める

$$w_{k,i,g} = \frac{\delta_{k,i,g}}{\sum_{g=1}^{G_k} \delta_{k,i,g}}$$

$$\delta_{k,i,g} = \exp\left(u_{k,i,g} + v_{k,i,g}\right)$$

  - $u_{k,i,g} \sim N\left(\bar{u}_{\delta_i}, \tau_g^{(u)}/n_{\delta_i}\right)$：ICARモデル（空間相関をもつ）
  - $v_{k,i,g} \sim N\left(0, \tau_g^{(v)}\right)$：各地域が独立
  - $(\tau_g^{(u)})^{-2} \sim U(0,500)$
  - $(\tau_g^{(v)})^{-2} \sim U(0,100)$

60

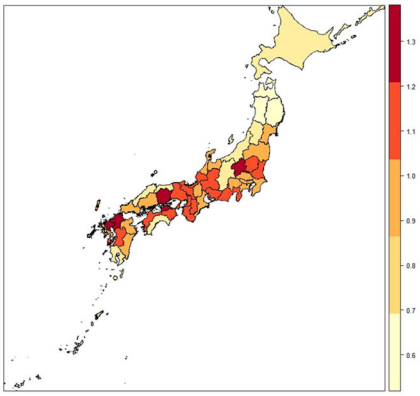
それから所属割合（重み）で、このようなモデルを考えております。

δというパラメータを、空間相関を持つよう、ICAR モデルに従う u と、地域ごとに独立な v の和からexp(u+v)という形で、正の値を取るように与えます。そのδはグループごとに異なる値を取っていて、δの全グループ計に対する各グループ g におけるδの割合を、所属割合にしましょうという形でモデル化しています。これを k が 1 の場合はクレーム頻度、k が 2 のときは保険金単価というように、同じ形のモデルで別々に推定しています。

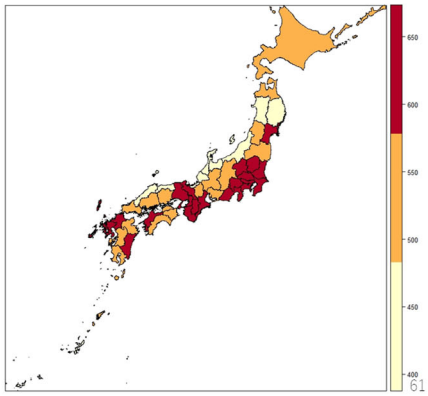
空間クラスタリング 自賠責保険での解析例

- 自家用乗用自動車の2016～2018年度計のデータへ適用
- 適度なグループ数にするためグループ間較差を$\beta = 0.2$、グループ数を頻度では$G_1 = 5$、単価では$G_2 = 2$とした

（全国を1とした）相対クレーム頻度



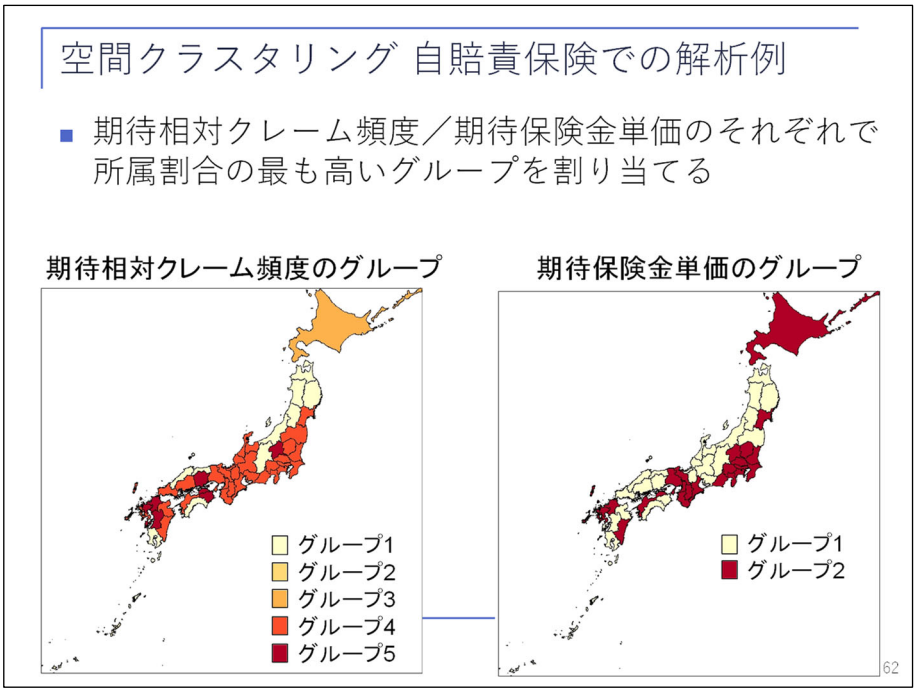
保険金単価（千円）



61

続いて解析例を紹介しますが、今度は時間的な変化は見ずに、2016～2018 年度計の自家用乗用自動車保険のデータへと適用しました。適度なグループ数にするために、先に全国を 1 とした場合のデータにおけ

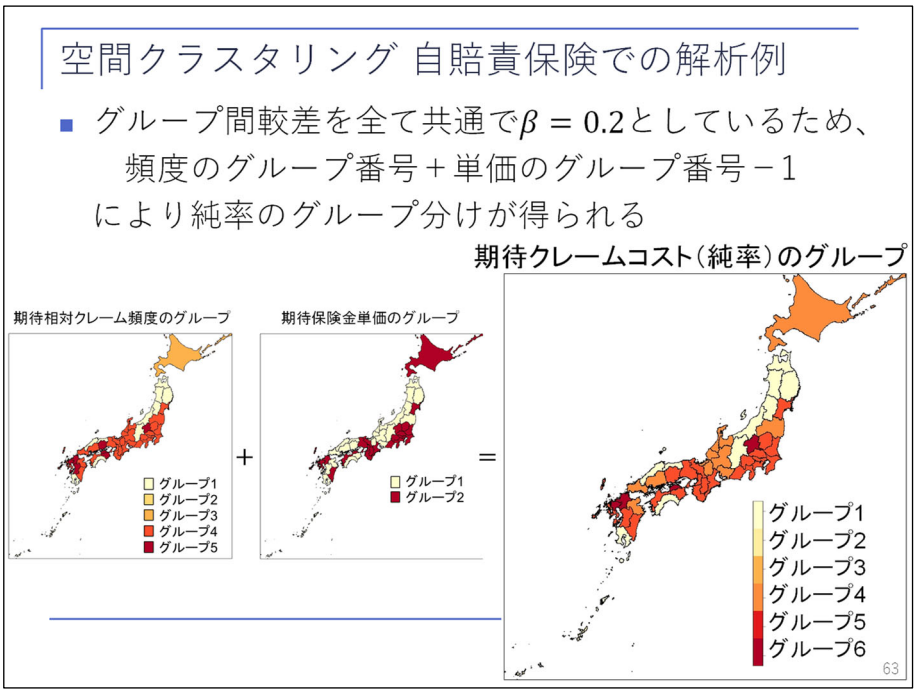
る相対クレーム頻度や、データにおける保険金単価の格差をチェックし、その結果、適度なグループ数にするためには、グループ間の較差を共通で $\beta = 0.2$ と、こちらで与えました。さらに、頻度のモデルにおいてはグループ数を 5 個、単価のモデルにおいてはグループ数を 2 個と設定しました。



この設定で各都道府県の所属するグループを推定した結果がこちらになります。

所属グループの推定は、推定された所属割合について、所属割合の最も高いグループをその都道府県のグループとして割り当てるようにしています。

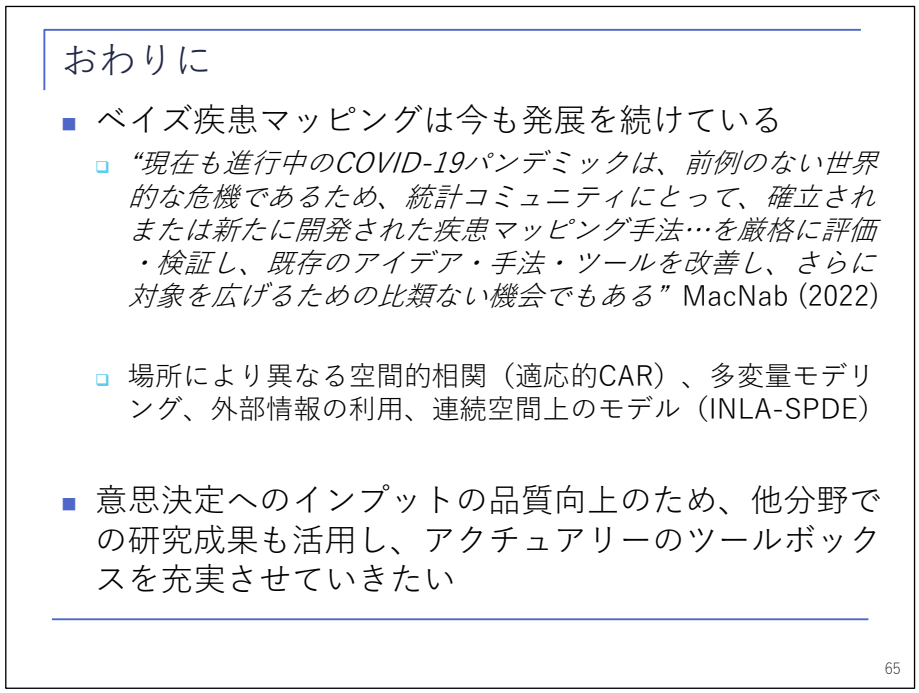
クレーム頻度の方は、グループ 1 から 5 にこのように分かれまして、保険金単価の方はグループ 1 と 2 に分かれます。



最後に、グループ間較差は全て共通にしていますので、頻度のグループ番号に単価のグループ番号を足して、最後に 1 を引くことによって、頻度と単価を合わせたクレームコスト、すなわち純率のグループ分けを作ることができます。

頻度が 5 段階、単価が 2 段階あって、合わせて 6 段階のグループになりまして、多少空間的な相関も持ちながら、グループが変化している様子がお分かりいただけるかと思います。

このように、都道府県を適度な数のグループ数に分けて、グループごとに料率を決める方法を提案いたしました。

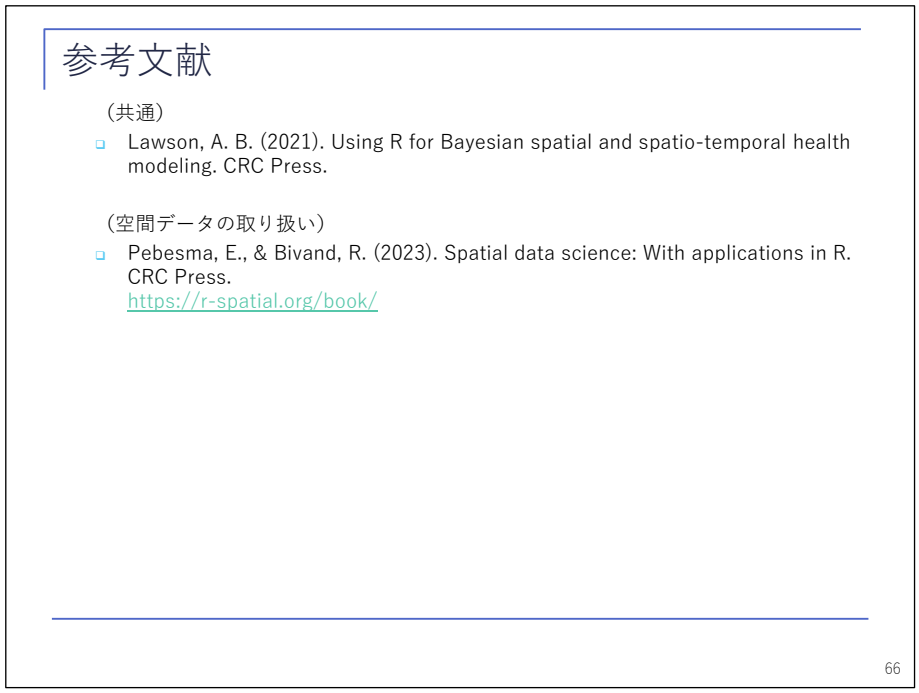


では、最後になりますが、このようなベイズ疾患マッピングは、今も発展を続けております。

特に現在も進行中の COVID-19 のパンデミックは、前例のない世界的な危機であり、統計コミュニティにとっても、確立されまたは新たに開発された疾患マッピングの手法を厳格に評価・検証し、既存のアイデア・手法・ツールを改善し、さらに対象を広げるための比類ない機会でもあると述べられています。

発展した手法ですと、場所により異なる相関を持たせるような適応的 CAR モデル、多変量モデリング、外部情報の利用、連続空間上のモデリングなど、いろいろと既に提案されております。

このようなモデルは、意思決定のインプットの品質向上につながるものでもございますので、アクチュアリーに限らず、他分野の研究成果も活用していきながら、アクチュアリーのツールボックスを充実させていければと考えています。



参考文献はこちらのとおりとなります。ご清聴いただきありがとうございました。

【司会】 それでは質疑応答に入ります。会場にいる方で質問のある方は挙手をお願いします。

【質問者】 貴重なお話をいただきまして、ありがとうございました。愛知県でこの区とこの区が隣り合っている、この市とこの市が隣り合っているということを表現されたと思いますが、市どうしであれば何かしらの重心めいたものどうしの距離を用いているのか、都道府県であれば県庁のある場所どうしの距離を用いて表現なさっているのか、そこはいろいろと余地があると思うのですが、今回のアメリカのものはどのようなものを用いられているのでしょうか。そこが気になった次第です。

【渡辺】 ありがとうございます。おっしゃるとおりいろいろな定義のしかたがあると思いますが、今回のテキストで使っているアメリカの例では、境界線が接しているものを隣接していると定義しています。

例えば、人口の重心間の距離にするのか、人が相互に行き来する量が多ければより密接に隣接しているなど、いろいろなやり方が考えられると思いますが、今回やったものは非常に単純な例です。

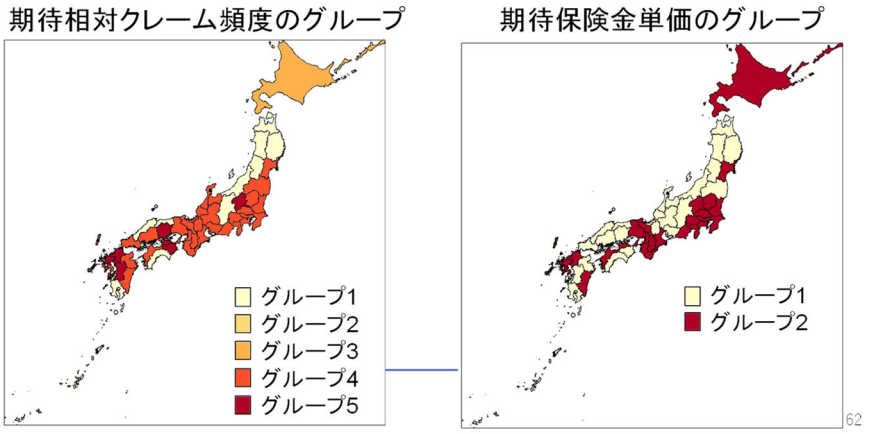
【質問者】 すみません、再び質問します。自賠責保険の結果で、これを定性的にどのように捉えればいいのかと思ったものがございまして、単価についての結果では都心部に近いほど色が濃い方にランクづけされていたと思います。ただ、頻度については高知県ではランク 1 だったけれども、隣の県だと急にランク 3 に跳ね上がるなど、隣の県なのに 2 ノッチも上がるということが観測されたかと思います。

実際に世に出ている地震保険や落雷であれば、隣の県で 2 ノッチも 3 ノッチも上がらないように調整されたりすると思うのですが、今回の自賠責の分析結果では、なぜ隣の県なのに 2 ノッチも上がったのか。データのばらつきなのかなど、そこがどのように解釈できるのかなと思った次第です。長々とすみません。

【渡辺】 ありがとうございます。野村さんの分析なので、私がどこまで答えられるか、どこまで答えていいのかはあるのですけれども、お答えできる範囲でお話しします。

空間クラスタリング 自賠責保険での解析例

- 期待相対クレーム頻度／期待保険金単価のそれぞれで所属割合の最も高いグループを割り当てる

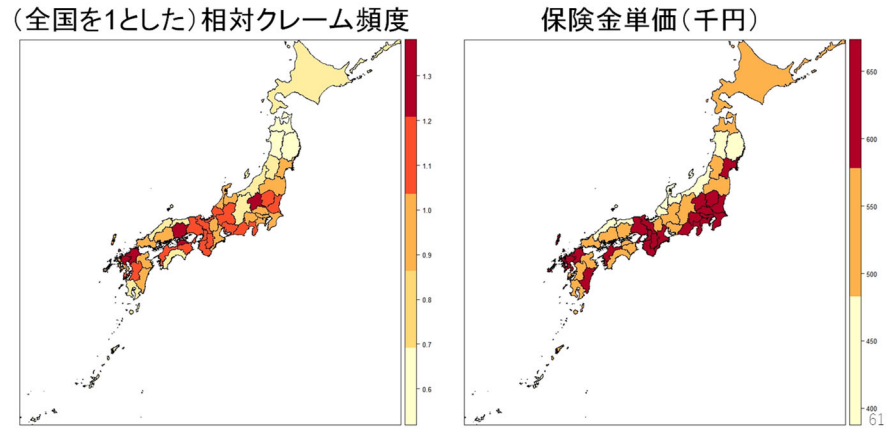


ここのことを言われていたのかなと思います。単価であれば、都市部の方がグループ 2 にあって高いということで、納得感はあるのですが、一方で頻度の方は、おっしゃるとおり不思議なことが起きているのではないか。高知県はグループ 1 だけれども、徳島県、香川県はグループ 5 になっている。

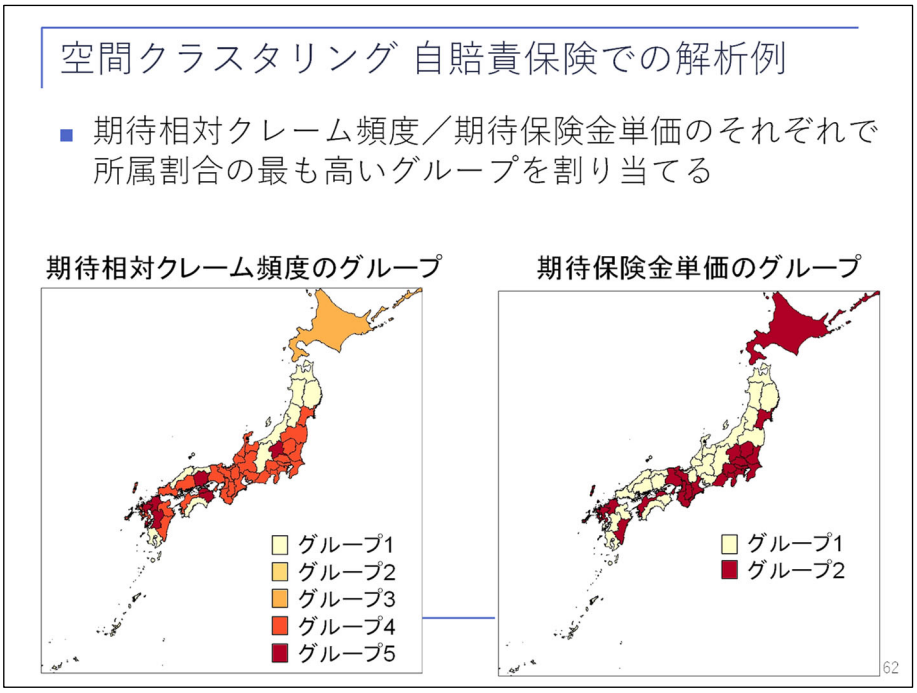
結論から言うと、私にもよく分かりません。分からないのがどのようなところから来るかというと、この一つ前に実績の相対クレーム頻度を出しているところがあると思います。

空間クラスタリング 自賠責保険での解析例

- 自家用乗用自動車の2016～2018年度計のデータへ適用
- 適度なグループ数にするためグループ間較差を$\beta = 0.2$、グループ数を頻度では$G_1 = 5$、単価では$G_2 = 2$とした



これを見ると、そこまでの較差はついていないのですね。徳島県は、それほど色が濃くなっていない状態ですので、クラスタリングをやった結果、



なぜこのような極端な差になってしまったのか、私もよく分かりません。ここはもう少し検討する必要があるのか、あるいは野村さんが解をお持ちかもしれませんが、そこはお答えできないところで、ご了承ください。

【質問者】　ありがとうございます。

【野村による追加説明】　当日私が出席できなかったために渡辺さんからご回答いただきましたが、ここで補足させていただきます。

本解析のモデルは、あるグループ g に対する所属割合のパラメータ δ_g について、隣接する地域間で近い値を取るように設計されています。その結果、同じグループに属する都道府県の大きなまとまりが幾つかできています。ところが、1 ノッチの差で隣接するグループ間では所属割合の関連性を何も持たせていないため、隣接地域間でグループが異なる場合の較差は、大きなグループのまとまりに引っ張られて 2 ノッチ、3 ノッチの差が現れたものと考えられます。本モデルは、都道府県の大きなまとまりを作ることに主眼を置いているため、隣接地域間の較差を抑えるためにはさらなるモデルの改変が必要となります。

【司会】　他にご質問がないようでしたら、Slido には質問は入ってきていませんので、以上をもちましてセッション A-5「空間時系列モデルのアクチュアリー業務への応用」を終了します。

発表者の方に、もう一度盛大な拍手をお願いします。