2012.11.30 Tokyo

PostGIS と R の連携による地理・環境データ分析

国府田 諭（埼玉大学環境科学研究センター研究員）
Kouda Satoshi

報告の内容

1. そもそもの課題

- なぜ PostgreSQL/PostGIS と R の連携に至ったか

2. GIS データを含む連携フロー

- PostGIS と R をつないでいる現在の手法

3. 実例紹介

- 日本の地域別 CO_2 排出量や気象データ等を用いて

そもそもの課題

- PostgreSQL/PostGIS と R の連携に至った背景、デスクトップアプリとは違うスクリプト中心の処理フローを構築した理由など。

ビッグデータ時代の、集計・分析上の課題

大量のデータ × 様々な組み合せ × 様々な分析手法 × 様々な計算方法とパラメータ

公開される社会統計データ・GISデータの着実な増加

▼

検討すべきデータの組み合わせ・集計方法などが飛躍的に増加

政府統計の総合窓口（総務省）

国土数値情報（国土交通省）

データ分析手法も多様化、高度化

大量のデータ × 様々な組み合せ × 様々な分析手法 × 様々な計算方法とパラメータ

コンピュータの性能向上と統計学の発展が相互作用し、新しい分析手法が増加。とくに空間分析で顕著

一つの分析手法でも、多様な選択がある

例「クラスター分析」

データを「似たものどうし」のグループに分ける、多変量解析の一手法

データ間の「距離」の算出方法
（ユークリッド距離、マハラビノス距離、等々）
×
データどうしの（非）類似度の算出手法
（ウォード法、重心法、メディアン法、等々）

メニュー操作のデスクトップ GIS に限界

大量のデータ × 様々な組み合せ × 様々な分析手法 × 様々な計算方法とパラメータ

- 使うデータを揃えるまでが大変
- エクセル等の作業用ファイル、シートが増え、管理不能に

- ソフトに備っている手法、設定しか使えない
- 多様な分析、設定を試したい。しかし同じデータに対して操作を繰り返すのは負担が大きい

- 最大の問題

「何をどう処理したのか」、後で使える記録がまず残らない

データベースと統計処理言語の連携を模索

大量のデータ × 様々な組み合せ × 様々な分析手法 × 様々な計算方法とパラメータ

データベースで管理

⇒ 地理情報も扱うなら PostGIS が定番

何らかの統計処理言語

⇒ オープンソースなら R

- グラフや地図プロットもできる
- 外部プログラムも実行できる
- 結果を PDF 等に出力できる

全体を R スクリプトで一括処理することに決定

PostgreSQL/PostGIS の利点

- PostGIS に豊富な空間処理関数が備っている
- GIS で面倒な「測地系」「投影系」の管理と変換が容易
- PL/PgSQL で、複雑な一連のデータ操作も記録、再利用できる
- 付属する pgAmin III でデータ全体を把握でき、R から呼び出すデータやユーザ定義関数の確認も容易

R の特徴

- もともと統計分析と結果の視覚化を主眼とする言語、ソフト。そのため「ベクトル」（一次元の配列）を中心とする独特のデータ構造をもつ
- 条件分岐、ファイル入出力、外部プログラム起動など、他のプログラミング言語と同等の機能もある
- 利用目的に応じて様々なパッケージを導入する使い方が主

R によるグラフ・地図等の PDF 出力例

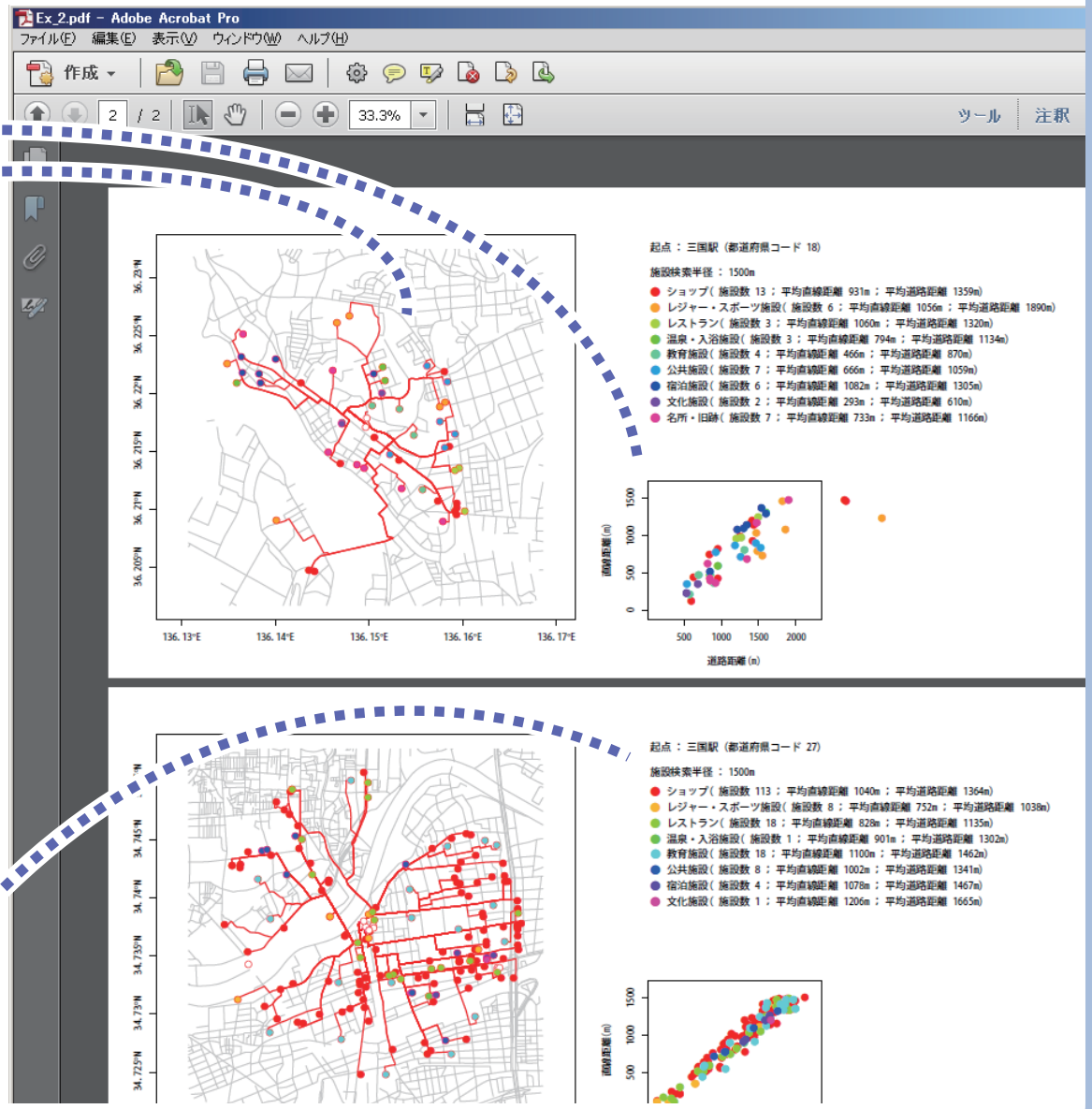
R には豊富な作図機能があり、地図・地理データの描画もグラフと同様に行える

作図結果を、様々なファイル形式で保存できる。

PDF への保存は、一枚の作図を一頁とし、一度に複数頁の PDF 出力ができる

日本語も問題なく出力できる。

（フォント埋め込みは現在のところ非対応だが、出力後に Acrobat で埋め込める）



スクリプト中心による処理の利点と目的

システム上の利点

- 全ての処理過程がテキストで記録されている
- 似た処理の複製が容易
- 複数スタッフ間の処理の共有、継承、カスタマイズが容易
- 処理の所要時間を記録でき、予測も容易
- 途中の処理ミスがあった場合、修正と再実行が容易

目的

- 手作業では不可能な量・質の分析を気軽に実行し、
  新たな事実発見や仮説構築につなげる
- マウスやメニュー画面操作の時間を減らし、
  人による思索と分析の時間を増やす

GIS データを含む連携フロー

- PostreSQL/PostGIS と R を連携させて
  現在実行しているフローの概要。
  また用いている R のパッケージなど。

現在の連携フロー（概要）

① 数値等のデータを読み込む
（R パッケージ：RPostgreSQL）

② 地理データを読み込む
（R パッケージ：rgeos）
※データ量が多い場合は別手法（次頁）

③ 数値等と地理データを R で結合、集計や分析

④ 結果をグラフや地図で PDF 等に出力
（R パッケージ：maptools）

⑤ 集計・分析結果を PostgreSQL/PostGIS に保存

全体を一つの R スクリプトで実行

PostgreSQL/PostGIS データベース

連携の要：R の三つのパッケージと pgsql2shp

RPostgreSQL

R から PostgreSQL に接続、クエリ発行、結果取得できる

※ Windows XP SP3、PostgreSQL 9.1.6、R 2.15.2 で動作確認。
以前、R 2.15.1 で試した時は接続できなかった

rgeos

WKT 形式（PostGIS が出力できる地理データ表現の一種）を読み込める。RPostgreSQL との併用で PostGIS と連携実現

※ データ量が多い場合はパフォーマンスが悪い

maptools

SHAPE 形式（代表的な地理データのファイル形式）を読み込める。データ量が多く rgeos で厳しい場合に有用。

※ 地理データの処理やプロットにも使う、定番的パッケージ

pgsql2shp

PostGIS 付属のコマンドラインツール。PostGIS の地理データを SHAPE 形式に書き出せる。

連携できるまでの模索の過程

PL / R

PostgreSQL 内のユーザ定義関数として R の機能を導入できるが

- 処理全体を制御するには不向き

ODBC

PostgreSQL 用ドライバ、R のパッケージ（RODBC）を使えば連携はできるが

- データベースごとにデータソースを設定し管理するのは手間だし柔軟性を欠く

RpgSQL

RPostgreSQL とは別の、PostgreSQL に接続するという R のパッケージだが

- ロード自体、成功したことがない

psql

RPostgreSQL が動作する前は、このコマンドラインツールを R から実行し、クエリの結果を標準出力またはファイル経由で R に読み込んでいた。

R から PostgreSQL への接続コード例

```
# パッケージロード。大文字・小文字を区別するのでスペリング注意
library(RPostgreSQL)
con = dbConnect(PostgreSQL(), user="user", password="***", dbname="***")

# クエリに 2 バイト文字を含む場合、事前に下記クエリが必要
dmy = dbSendQuery(con, "SET client_encoding = '***'")

sql = "
    SELECT * FROM t_menbers WHERE f_name IN (
          ' 長谷川 ', ' 三日月 ', ' 柏崎 ', ' 高山 ', ' 楠 ', ' 志熊 '
    ) ; "
# SQL を複数行に分けて記述でき、字下げも可能

res = dbGetQuery(con, sql)
# これでクエリ結果が「データフレーム型」の変数 res に入った

# 次頁に続く
```

なお R での変数への代入は、= だけでなく <- も使える

R から PostGIS データを取得する例（1）

```
# 前頁から続く

# パッケージロード
library(rgeos)

# クエリに 2 バイト文字を含む場合、事前に下記クエリが必要
# dmy = dbSendQuery(con, "SET client_encoding = '***'")

res = dbGetQuery(con, "SELECT gid, ST_AsText(geom) FROM t_polygons")
sps = readWKT(res$st_astext[1])
# これで t_polygons の最初の地物が、地理データとして変数 sps に入った

plot(sps)
# 変数 sps の地理データがプロットされた

dbDisconnect(con)
```

PostGIS の ST_AsText() 関数で、地物を「WKT 形式」で出力。
これを rgeos パッケージの readWKT() 関数で一行ずつ読み込む。

R から PostGIS データを取得する例（2）

```
# パッケージロード
library(maptools)

# pgsql2shp を起動するコマンド文字列を作成
cmd = '***/pgsql2shp.exe -u *** -P *** -f "shape_path" dbname tablename'

# pgsql2shp を起動
system(cmd)
# これで shape_path に SHAPE 形式のファイルが作成され、PostGIS のテーブル
tablename の内容が格納された

map = readShapeSpatial("shp_path")
# これで SHAPE の内容全体が、地理データとして変数 map に入った

plot(map)
# 変数 map の地理データがプロットされた
```

PostGIS 付属の pgsql2shp.exe を R から起動し SHAPE を出力。
これを maptools パッケージの readShapeSpatial() 関数で読み込む。

実例紹介

① データの視覚化（塗り分け地図等）

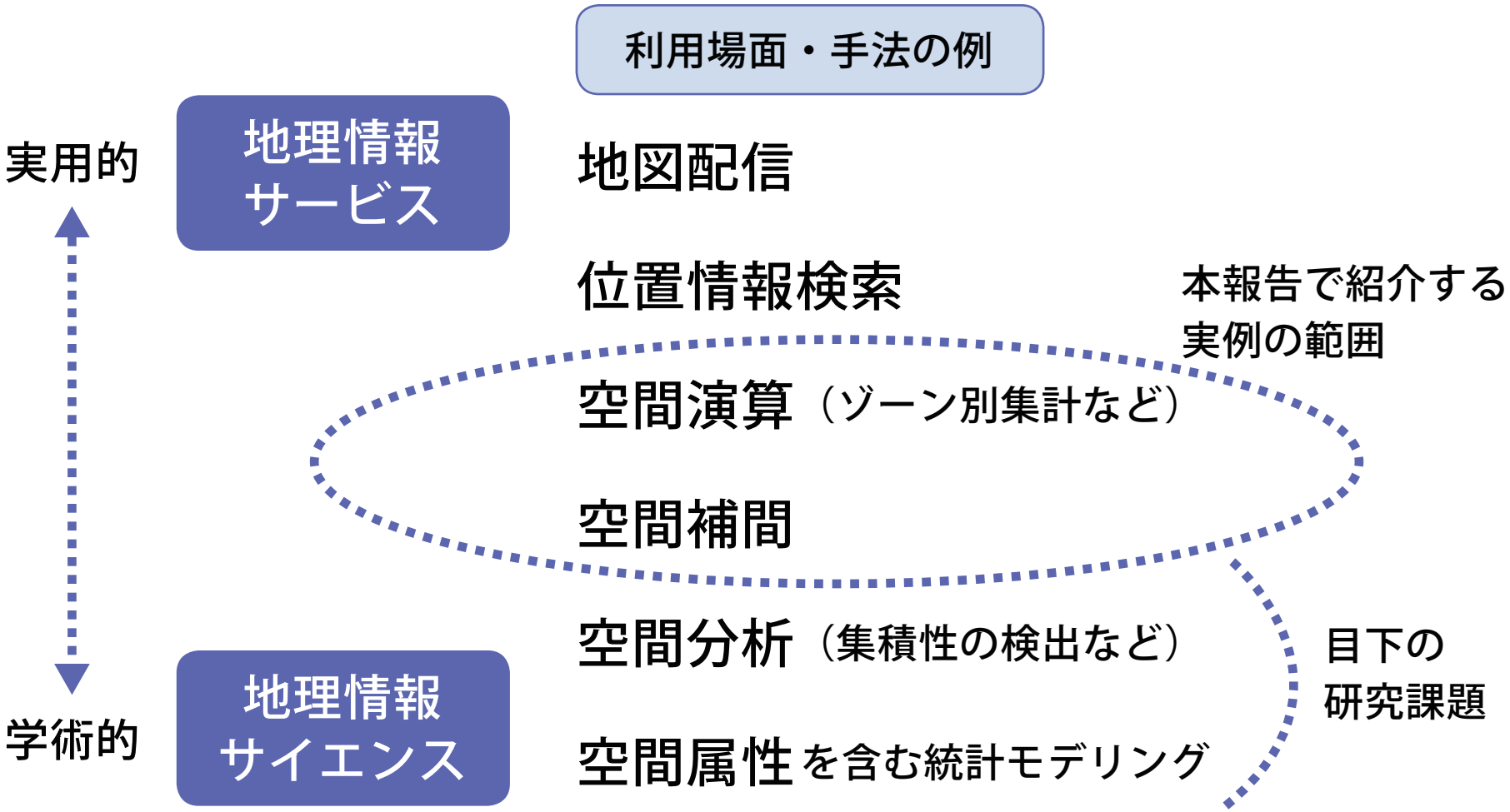
② データの空間補間

③ 距離帯別のデータ集計と分析

紹介する図は、全て PostgreSQL/PostGIS と R で PDF に出力したままの状態です。
カラー部分が多いですが、白黒では違いが不明瞭な場合もあります。ご容赦ください。
作成環境は、Windows XP SP3, PostgreSQL 9.1.6, PostGIS 2.0.1, R 2.15.2 です。

GIS の主な利用分野と「二つの S」

GIS は「地理情報システム」の略語だが、利用目的・分野によって
S に「サービス」または「サイエンス」が含意されることもある



実例① 市区町村別 CO_2 排出データの視覚化

データ： 環境省「地球温暖化対策地方公共団体
実行計画（区域施設編）策定支援サイト」から

「部門別 CO_2 排出量の現況推計」

http://www.env.go.jp/policy/local_keikaku/kuiki/tools.html

「ドットマップ」

対象地域が密集していると
点が重なり、見にくい

「コロプレス図」

いわゆる塗り分け地図。
データを色で分けて表現

部門別CO2排出量の現況推計

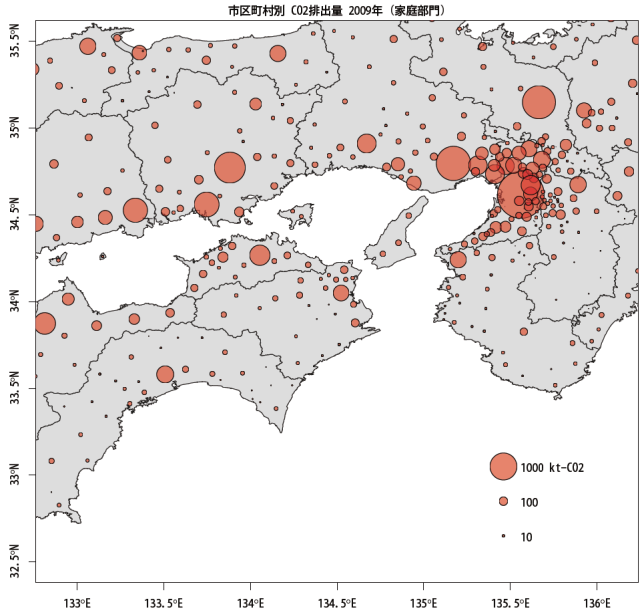
簡易版マニュアルに基づき全市区町村の部門別CO2排出量の現況推計値を算出しました。下表からダウンロードして御活用ください。（推計結果は飽くまでも参考値としてご活用下さい）

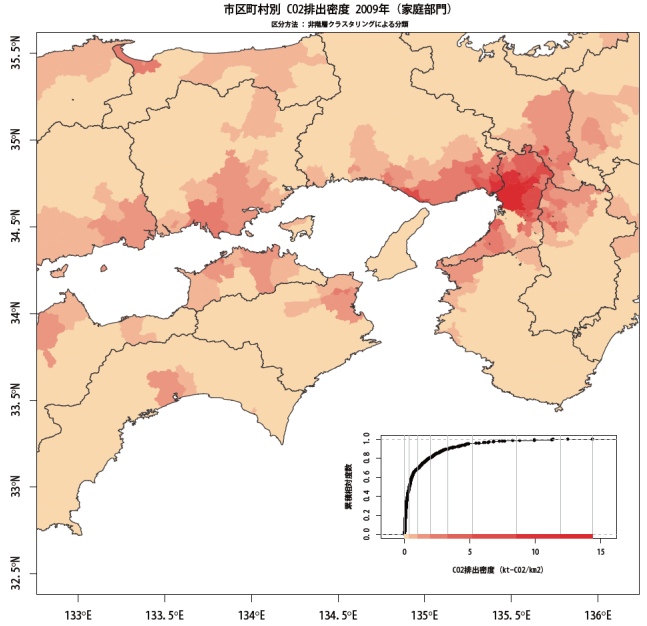
排出量エクセルデータ一覧

部門		1990年度	2007年度	2008年度	2009年度
産業部門	製造業	ダウンロード [XLS:763KB]	ダウンロード [XLS:396KB]	ダウンロード [XLS:355KB]	ダウンロード [XLS:382KB]
	建設業・鉱業	ダウンロード [XLS:830KB]	ダウンロード [XLS:373KB]	ダウンロード [XLS:405KB]	ダウンロード [XLS:375KB]
	農林水産業	ダウンロード [XLS:831KB]	ダウンロード [XLS:373KB]	ダウンロード [XLS:405KB]	ダウンロード [XLS:374KB]
家庭部門		ダウンロード [XLS:940KB]	ダウンロード [XLS:376KB]	ダウンロード [XLS:946KB]	ダウンロード [XLS:375KB]
業務部門		ダウンロード [XLS:824KB]	ダウンロード [XLS:374KB]	ダウンロード [XLS:405KB]	ダウンロード [XLS:376KB]
運輸部門	自動車	ダウンロード [XLS:1.01MB]	ダウンロード [XLS:1.05MB]	ダウンロード [XLS:906KB]	ダウンロード [XLS:918KB]
	鉄道	ダウンロード [XLS:916KB]	ダウンロード [XLS:373KB]	ダウンロード [XLS:344KB]	ダウンロード [XLS:354KB]
	船舶	ダウンロード [XLS:230KB]	ダウンロード [XLS:393KB]	ダウンロード [XLS:363KB]	ダウンロード [XLS:367KB]
廃棄物部門		ダウンロード [XLS:730KB]	ダウンロード [XLS:2.03MB]	ダウンロード [XLS:909KB]	ダウンロード [XLS:767KB]
合計		ダウンロード [XLS:2.82MB]	ダウンロード [XLS:2.81MB]	ダウンロード [XLS:3.47MB]	ダウンロード [XLS:3.39MB]

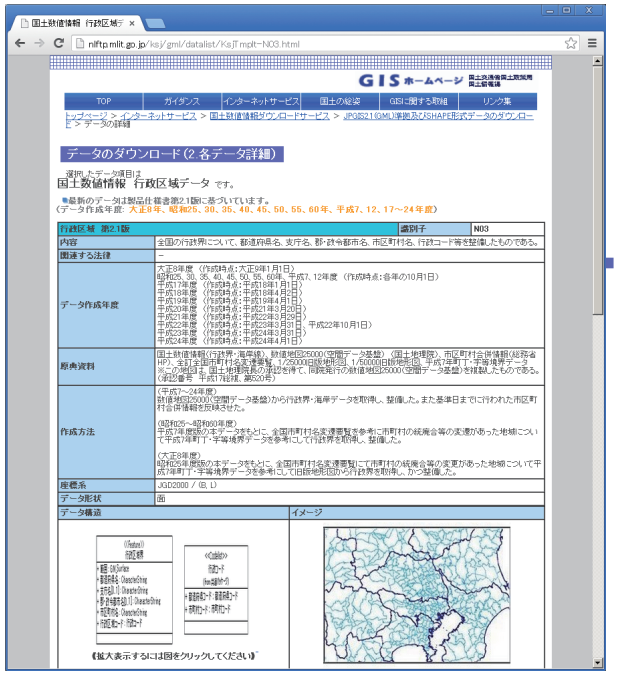
注1 全体：都道府県別エネルギー消費統計については、1990年度及び2007年度については2009年度発表資料(暫定値)、2008年度は2010年度発表資料(暫定値)、2009年度は2011年度発表資料(暫定値)を使用している。都道府県エネルギー統計は、毎年見直しが行われるため、現在HPで公表されている数値とは異なる可能性がある。

注2 製造業：工業統計調査における製造品出荷額等のうち秘匿数値(x)については、「0」として入力を行った(秘匿数値(x)とは、1又は2の事業所に関する数値で、これをそのまま掲げると個々の申告者の秘密が漏れるおそれがあるため秘匿した箇所であり、3以上の事業所に関する数値であっても、1又は2の事業所の数値が前後の関係から判明する箇所も秘匿としている。





いわゆる塗り分け地図の作成手順（概要）



行政界データ： 国土交通省「国土数値情報」から
行政区域データ（JPGIS2.1 準拠及び
SHAPE 形式）をダウンロード

http://nlftp.mlit.go.jp/ksj/gml/datalist/KsjTmplt-N03.html

そのまま R で SHAPE を読み込んでもよいが、
ファイルが年別・県別に分かれており面倒。
まず PostGIS に取り込み、使いやすく整理する。

（今回は、1 年ずつ 1 テーブルに統合）

① RPostgreSQL パッケージで市区町村コードと数値データのセットを SQL で呼び出し、データフレームに入れる

▶

② PostGIS 付属の pgsql2shp で、市区町村コードと地図データのセットを SHAPE に出力

▶

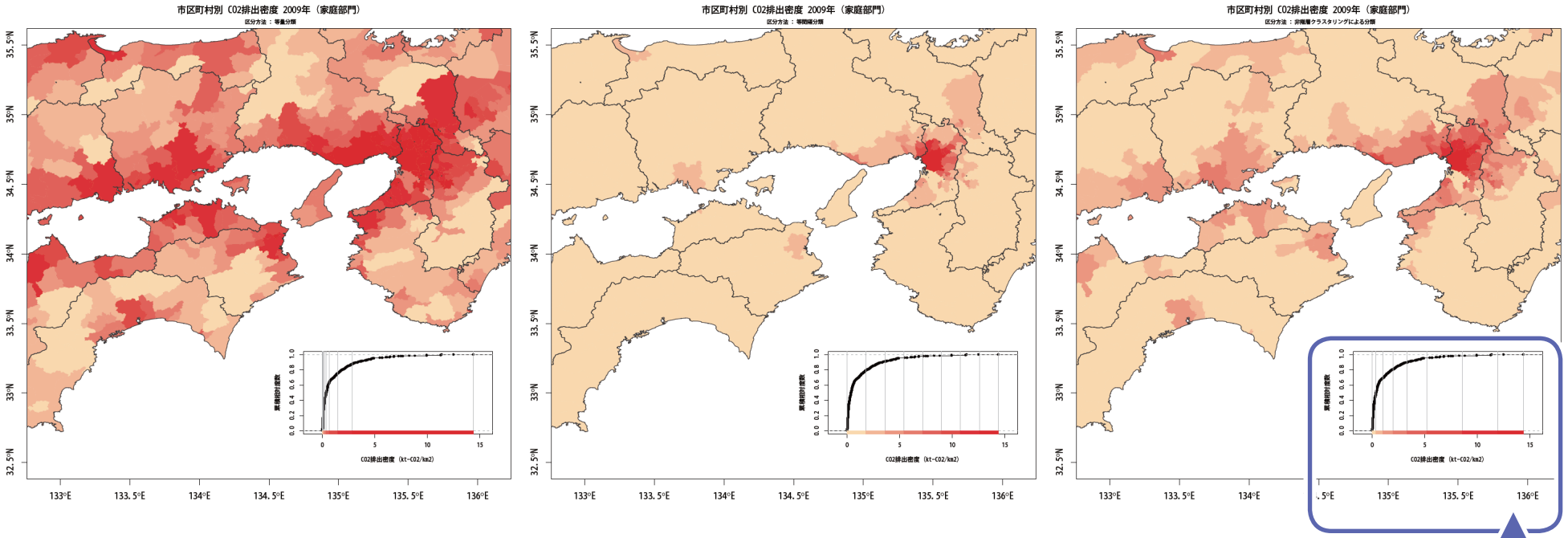
③ R の maptools パッケージで SHAPE を読み込む。市区町村コードをキーに数値データと結合し、PDF 等に出力

色分けの階級区分法によって大きな差

「等量分類法」
各色が同じ数になるよう区分

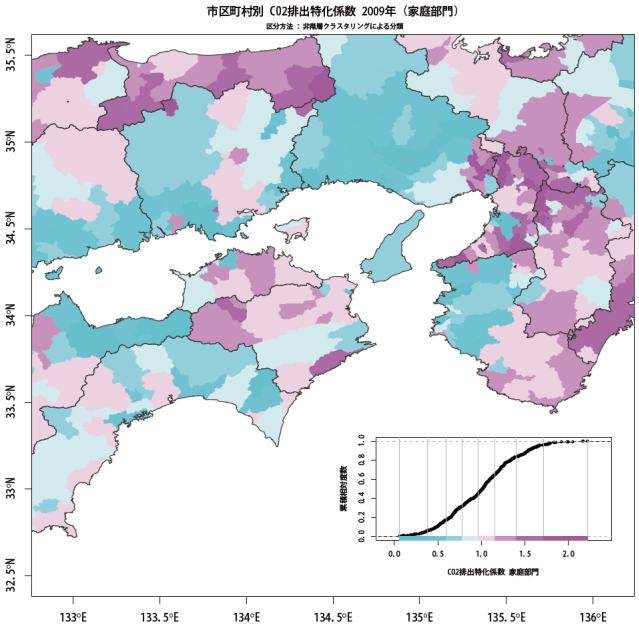
「等間隔分類法」
数値を等間隔に区分

「非階層クラスタリング」
による分類法

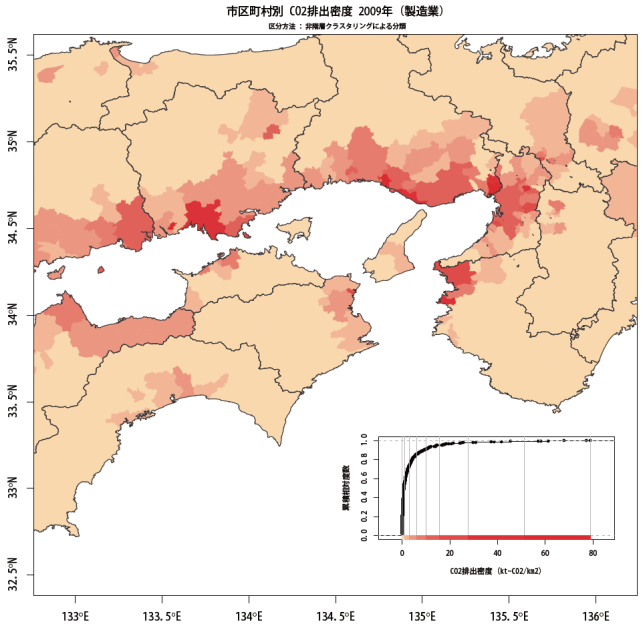


- 三つとも同じデータ。色分けの階級区分法でこれだけ差が出る。
- 等量分類法は、数値に大差のある地域が、不自然に同じ色になる可能性が多い。
- この他にも多様な区分法がある。R のパッケージ（classInt）では 9 種類可能
- 最良の方法や「定石」はない。実データの分布を合わせて示すのが良いと思う。

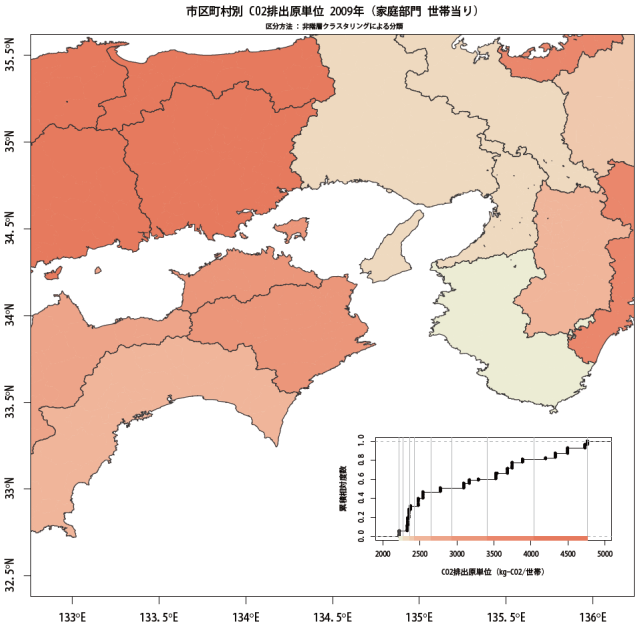
多様なデータ、指標で連続作成



↙ 地理学でよく使われる「特化係数」という指標を、
家庭からの CO_2 排出について算出。
赤系統が強い地域ほど、家庭からの排出割合が多い

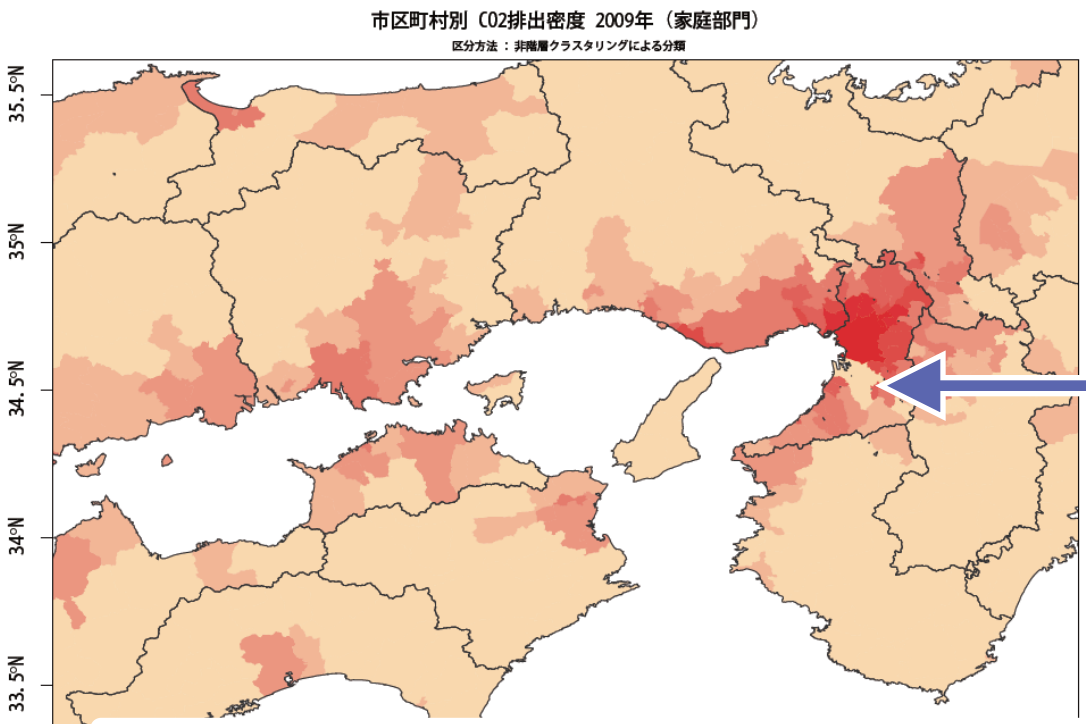


↙ 製造業からの CO_2 排出。
瀬戸内の臨海部に多い



家庭でのエネルギー効率を示す「排出原単位」
を算出したら、府県ごとに同一値だった…
元データの内部構造が視覚的に分かる。→

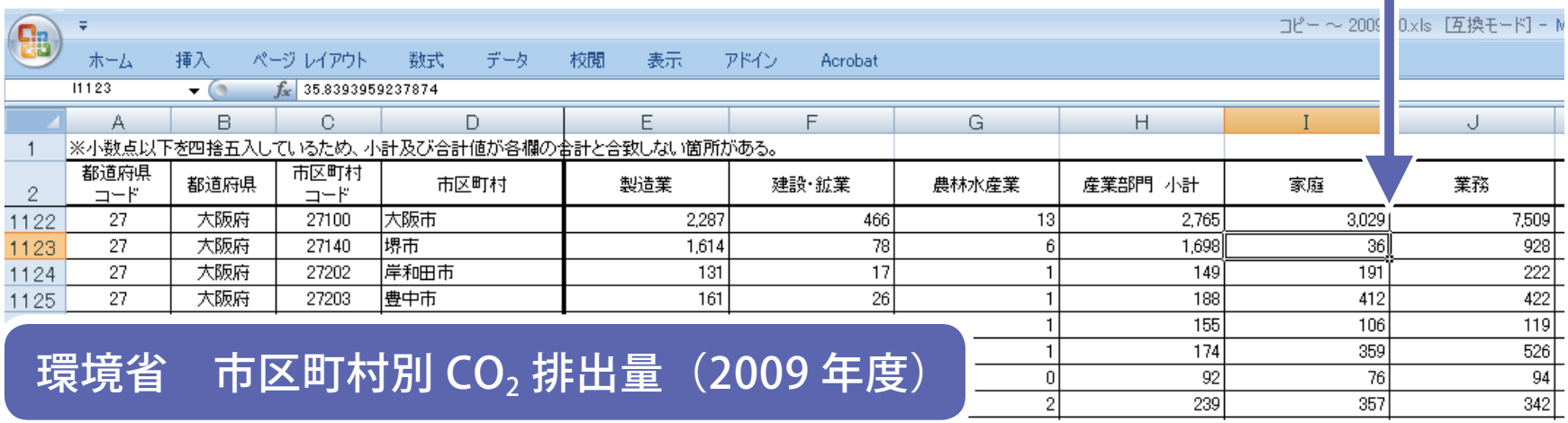
データチェックにも有効



大阪府の南部に、
周辺と比べて
排出量が目立って少ない地域

堺市あたり…
地図作成のミス？

元データ自体がそうなっていた。
周辺の市の数値と比較して、
明らかに少な過ぎる。
元データのミスと思われる。



※小数点以下を四捨五入しているため、小計及び合計値が各欄の合計と合致しない箇所がある。

都道府県コード	都道府県	市区町村コード	市区町村	製造業	建設・鉱業	農林水産業	産業部門 小計	家庭	業務
27	大阪府	27100	大阪市	2,287	466	13	2,765	3,029	7,509
27	大阪府	27140	堺市	1,614	78	6	1,698	36	928
27	大阪府	27202	岸和田市	131	17	1	149	191	222
27	大阪府	27203	豊中市	161	26	1	188	412	422
						1	155	106	119
						1	174	359	526
						0	92	76	94
						2	239	357	342

環境省　市区町村別 CO_2 排出量（2009 年度）

実例② 気象データの空間補間

気象データ：

気象庁「過去の気象データ検索」から
全国の気象台・アメダスの日別気温を
を閲覧・保存

http://www.data.jma.go.jp/obd/stats/etrn/index.php

気象台・アメダスの経緯度も得られる

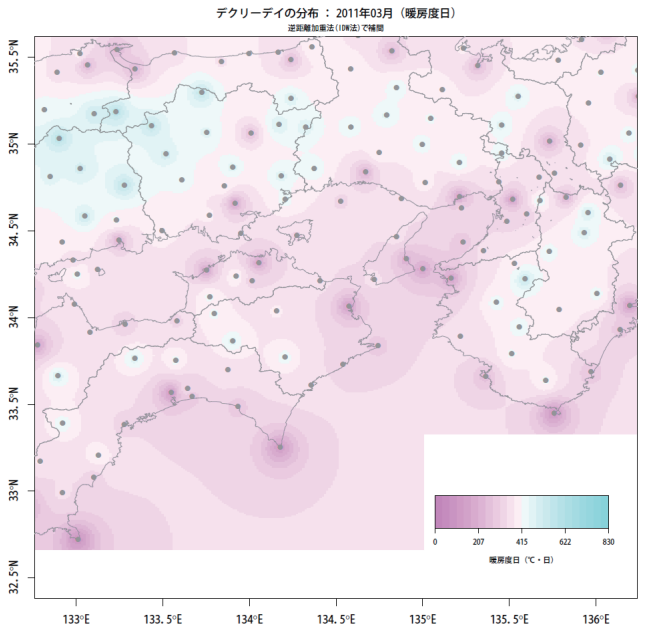
気象台・アメダスの日別気温

▼

エネルギー分野でよく使われる「デクリーデイ」
という、暑さ・寒さを示す指標を算出

▼

全国のメッシュ（約 1km 四方）で空間補間

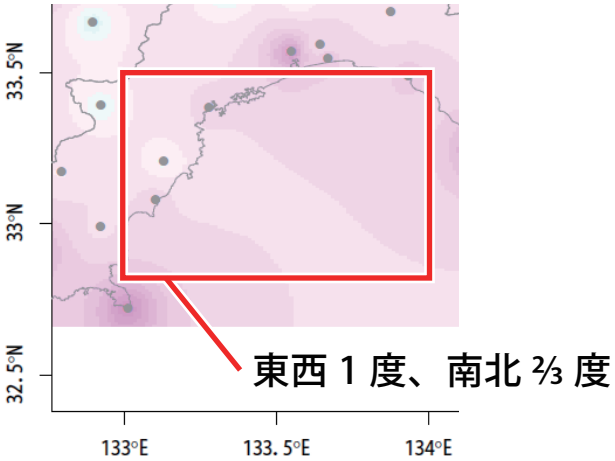


空間補間の手順（概要）

① PostgreSQL で、気象台・アメダスごと、年月ごとの「デグリーデイ」を算出、テーブルに保存

② R の RPostgreSQL パッケージで数値と気象台・アメダス座標データを読み込む

③ R で、補間したい地点の座標行列を作成
（今回は全国の 1km メッシュの各グリッド）

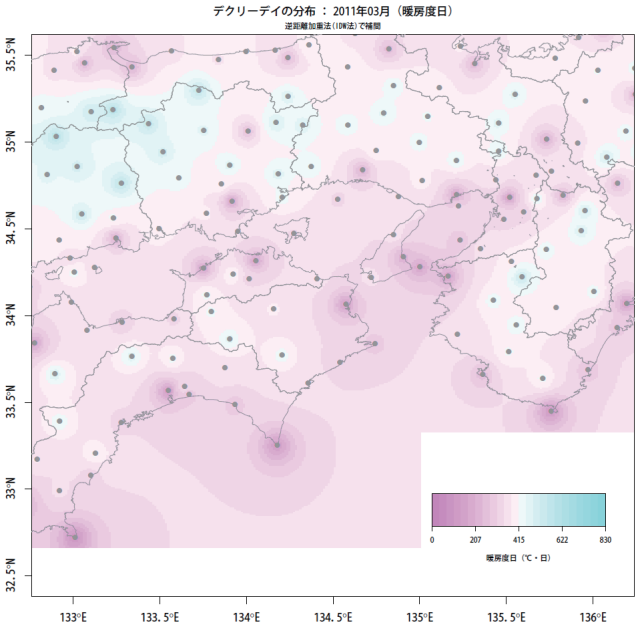


メッシュ … 経緯度などで地表をグリッド状に分割した区画。
上の範囲を縦横 80 等分すると約 1km 四方のメッシュになる

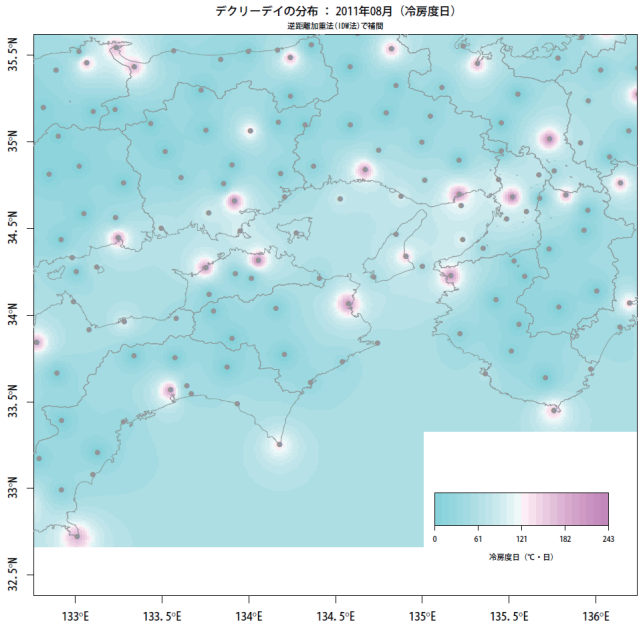
④ R の gstat パッケージに、②のデータと③の座標行列を渡し、補間方法を指定して実行するだけ。
（今回は、簡易な「逆距離加重法」を利用）

⑤ R で、補間結果データをカラー画像としてプロット

補間結果の例

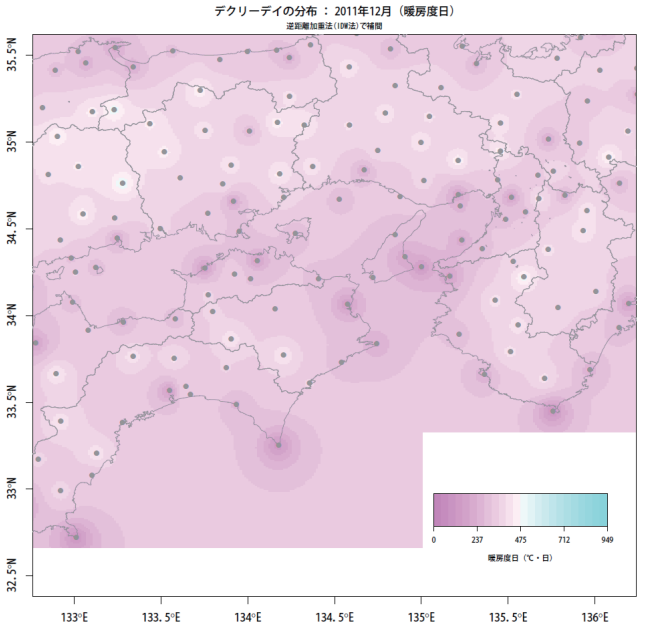


↙ 近畿・中国・四国地方、2011 年 3 月。
太平洋側や京阪神は比較的温暖（暖房度日が少ない）。
一方で岡山県・広島県の中山間部はまだ寒い。

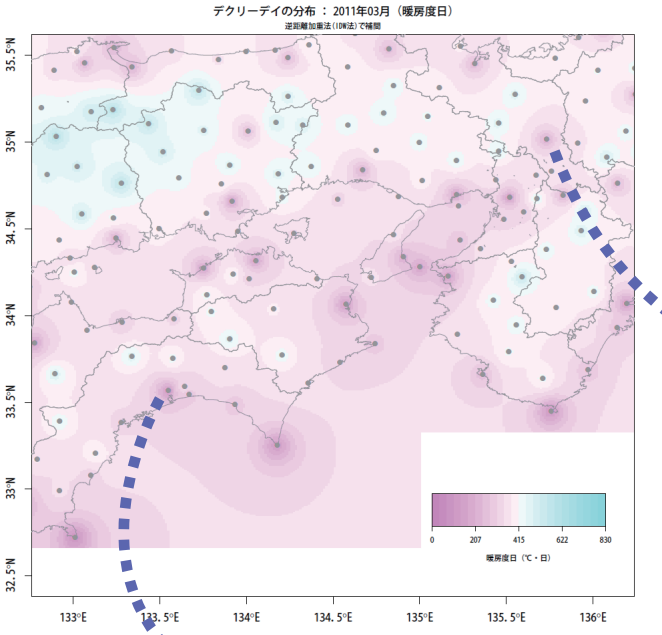


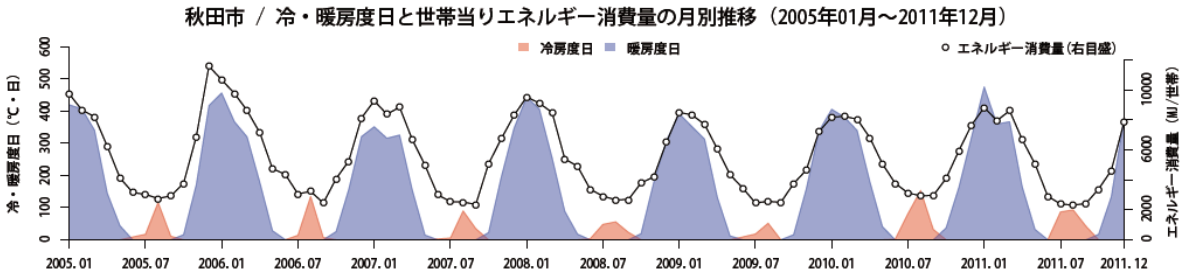
↙ 2011 年 8 月。暑さが
厳しいのは主に沿岸部。
内陸部は、京都市を除き
比較的涼しい。

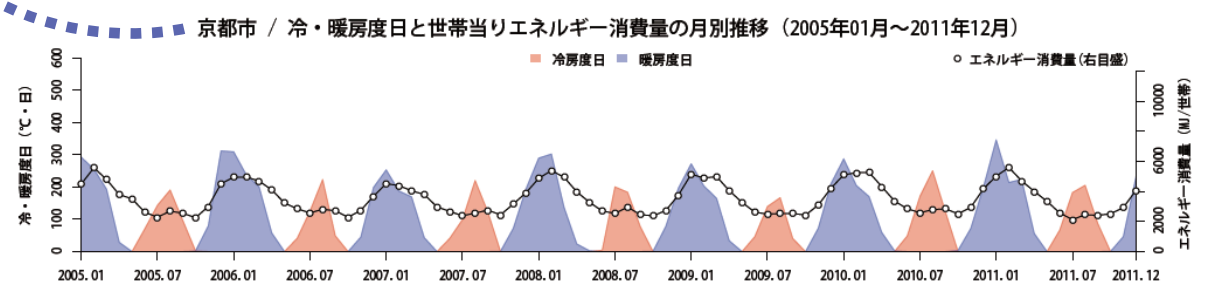
2011 年 12 月。この月の前半が全国的に暖かったため、同年 3 月よりも暖房度日が少ない。また地域的な差も少ない。→

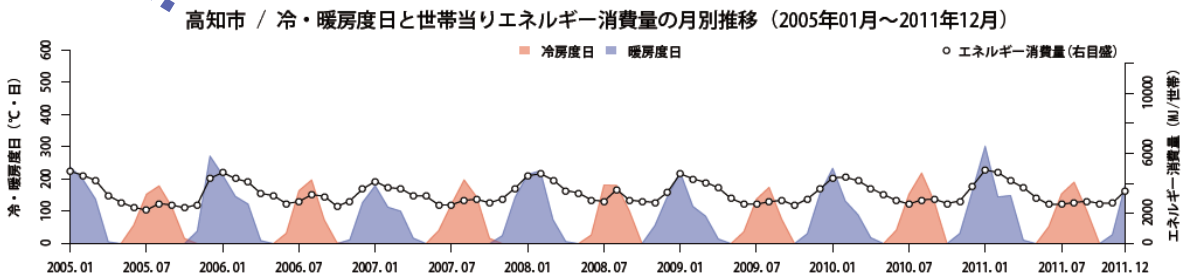


時系列変化を、エネルギー消費量と比較









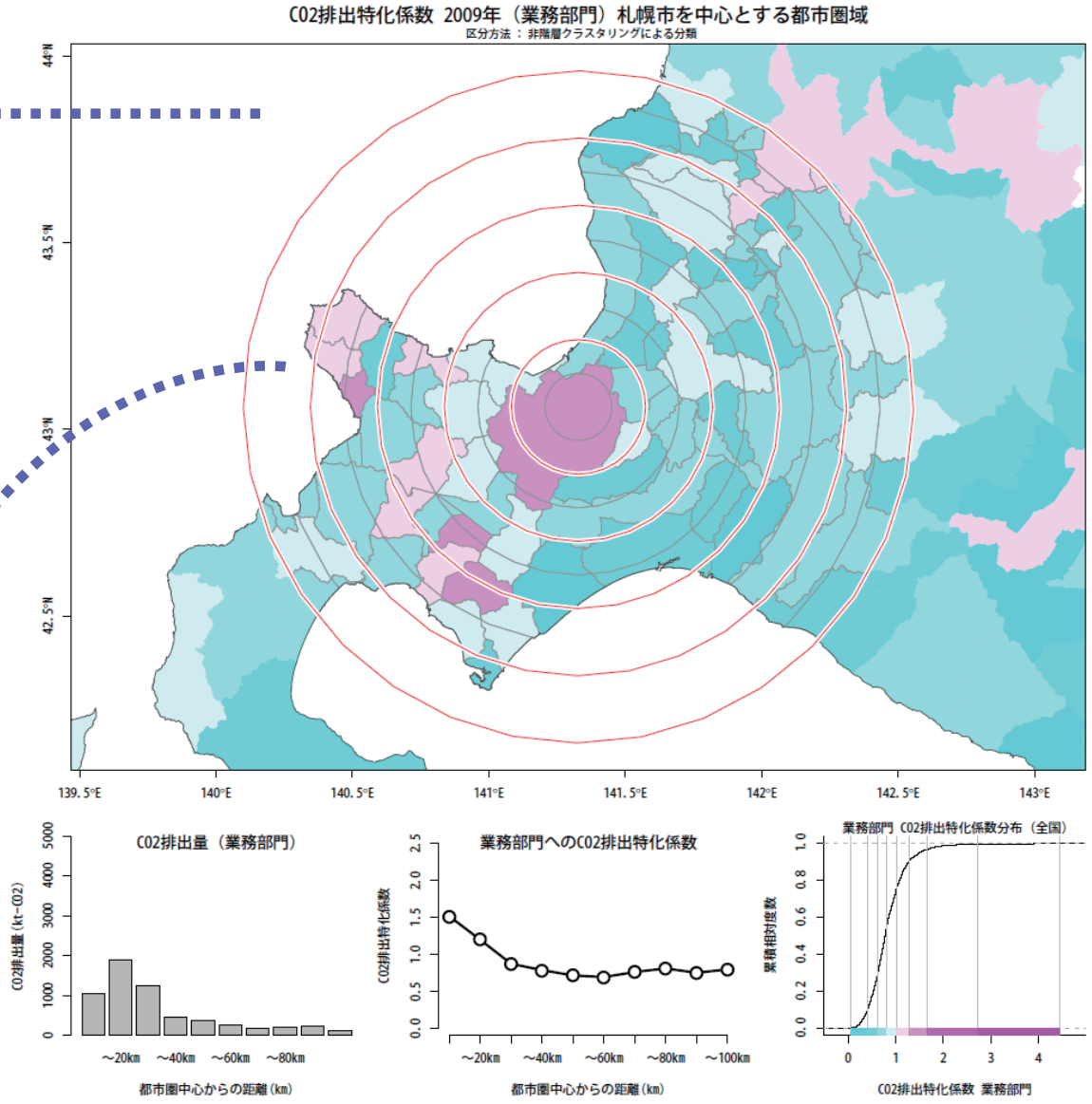
右上から秋田市、京都市、高知市。2005 年 1 月から 2011 年 12 月までの推移。

- 寒色と暖色の山 … デグリーデー。折れ線 … 世帯あたりエネルギー消費量。
- エネルギー消費量は家計調査月報（総務省）から集計した都市別データ。

実例③ 距離帯別のデータ集計と分析

実例①と同様、市区町村別の塗り分け地図を作成。
データは「業務部門」からの CO_2 排出特化係数。
オフィス、商業施設などの排出割合の多さを示す。

札幌市を中心に 10km ごとの距離帯を作成。
市区町村別データを距離帯別データに再集計。GIS でよく使われる「面積按分」という手法を用いる。
ともに PostGIS で実行。



距離帯別の結果を R でグラフ化。
都市圏ごとの比較分析を行う。→

GIS の集計でよく使う「面積按分」と PostGIS

すでに数値データが付いている地域区分 Ⓐ
（ここでは市区町村ごとの行政界）

新たにデータを集計したい地域区分 Ⓑ
（ここでは都市圏ごとの距離帯）

tb1
id1 geom1 num1

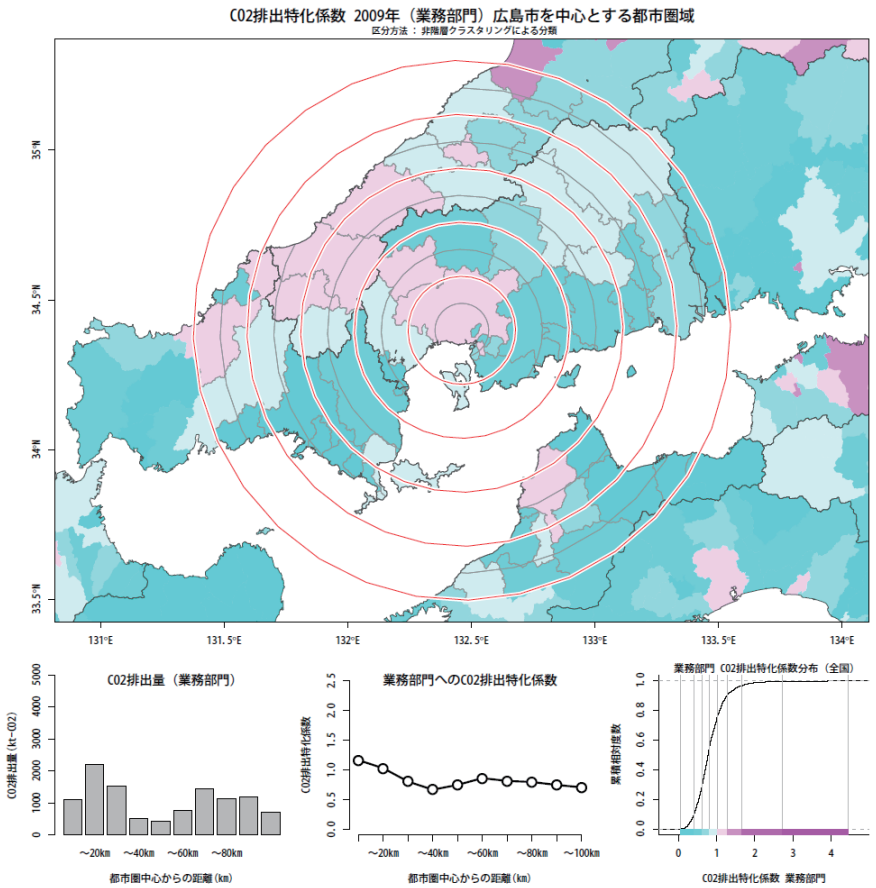
面積按分とは、地域区分Ⓐと Ⓑの
交差部分の面積の比率を使い、
地域区分Ⓑでの値を作成する手法。

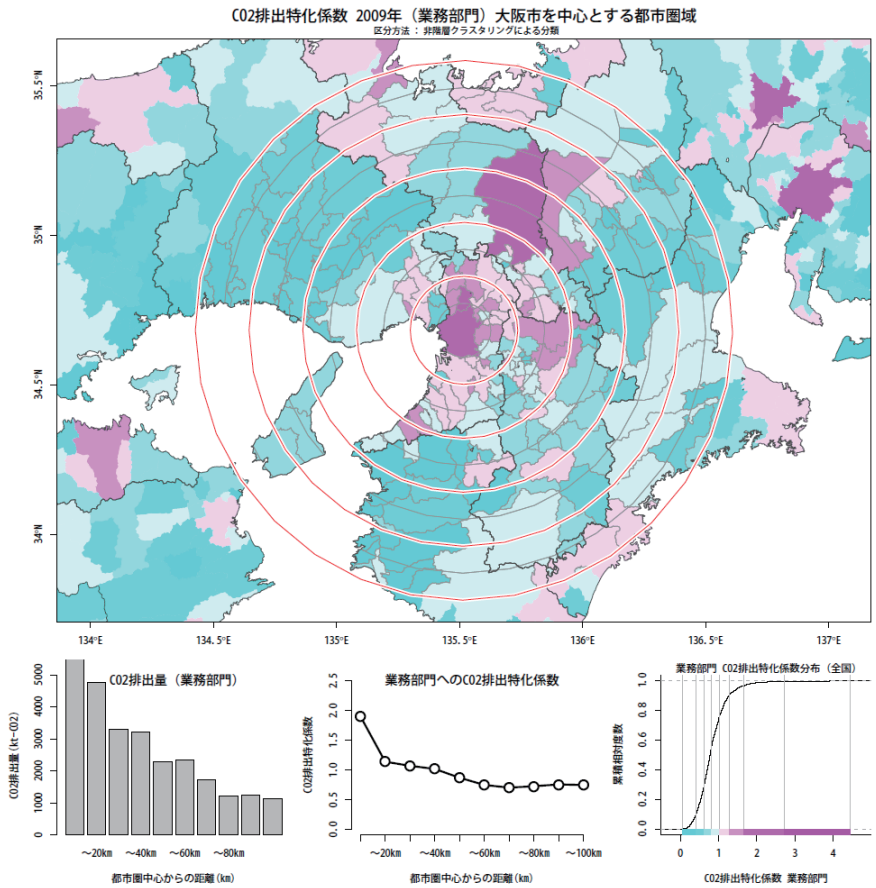
tb2
id2 geom2

PostGIS にテーブルがあれば、クエリー一つで算出できる

```
SELECT id2, sum(
      num1 * ST_Area(ST_Intersection(geom1, geom2)) / AT_Area(geom1)
) FROM tb1, tb2 WHER ST_Intersects(geom1, geom2)
GROUP BY id2 ;
```

大都市圏の比較（業務部門からの CO_2 排出特化係数）





- 広島都市圏（左）、京阪神都市圏（右）について集計結果を出力。
- 広島市は、都市圏中心市の割には特化係数が低い。製造業の方が多いため。
- 排出量（地図左下の棒グラフ）では、広島中心部より京阪神郊外の方が多い。

本資料で割愛した内容

- 実例における所要時間のパフォーマンス指標。用いた市区町村界データが過度に高精細なため、PostGIS による簡素化を図っているところです。その結果をもって発表時に報告します。
- 専門用語の解説。SHAPE 形式、メッシュ、デグリーデイ、特化係数、面積按分など、本資料では説明不足の点につき、発表時に補足します。
- PostgreSQL/PostGIS と R の連携に関する、今後の期待と展望。とくに PostgreSQL9.2 で導入された json 型、PostGIS2.0 で正式導入されたラスタ型には多くの活用可能性があると考え、検討を進めています。発表時に補足します。

▶ 本報告・資料に関する連絡先
国府田 諭 / satkouda@gmail.com