データ分析を社会のシミュレーションに利用する

佐藤彰洋
京都大学大学院情報学研究科
科学技術振興機構さきがけ

京都大学
KYOTO UNIVERSITY

2017年12月20日 計算機科学が拓く世界（後期 12回目）

さきがけ
PRESTO

自己紹介

- 佐藤 彰洋（さとう あきひろ）
- 大学院情報学研究科　特定准教授
- URL: http://ssuopt.amp.i.kyoto-u.ac.jp/akihiro/
- E-mail: aki@i.kyoto-u.ac.jp

授業の内容

1. はじめに
2. 社会のモデルの２面性
3. 社会のデータの集め方とシミュレーションの進め方
4. 社会のシミュレーションを行うための準備としてのデータ分析（事例）
5. まとめ

社会シミュレーション 世界を「見える化」する

横幹〈知の統合〉シリーズ編集委員会（編）

委員長　遠藤 薫

東京電機大学出版局

A5判　130頁　並製　（1,800円＋税）

ISBN　978-4-501-63070-6　C3000

奥付の初版発行年月　2017年09月

書店発売日　2017年09月20日

目次

第3章　データ分析を社会のシミュレーションに利用する（佐藤 彰洋）

1. はじめに
2. 社会のモデルの2面性
3. 社会のデータの集め方とシミュレーションの進め方
4. 社会のシミュレーションを行うための準備としてのデータ分析
5. まとめ

モデルの目的 (Joshua Epstein)

- 予測
- 説明
- 中心的な力学に焦点を当てる
- 動的なアナロジーを提案する
- 新しい疑問を発見する
- 科学的な精神的態度を喚起する
- もっともらしい範囲に成果を固定する
- 中心的な不確実性に焦点を当てる
- 実時間に近い時間内で危機の候補を挙げる
- トレードオフを実演する/効率性を提案する
- 外乱を通じて確立された理論の頑健性を試す
- 利用可能なデータと不一致となる確立された知恵を判明させる
- 実務家を訓練する
- 政策的対話を規律する
- 一般大衆を教育する
- 一件単純（複雑）な事柄が複雑（単純）であることを明らかにする

社会をシミュレートする目的

(1) **説明**

すでに起こっている社会現象を真似するモデルを作り、そのメカニズムを理解するためのシミュレーション。起こっている現象からあり得るストーリーを作りモデルとして表現する

(2) **推定・予測**

これから将来に起こる可能性のある現象について、既知のモデルと同じメカニズムに従うと仮定することで、観測できない部分の推定または将来の予測を行う

(3) **設計**

また存在しない社会システムを設計するときに、どのような社会を作るとどのようなことが起こり得るかを考察するためにシミュレーションする場合がある。この場合の目的は、設計の仕様とメカニズムをどのように社会に組み込むのかを問うことにある

シミュレーションの目的と方法、内容

（1）**説明**　現状の観測と一致するモデルの選択が必要

（2）**予測**　現状の観測から理解される状況が将来においても継続していることを仮定することが必要→時間的斉一性(uniformity)

（2）**推定**　観測されていない部分においても観測できる部分とメカニズムが連続していることを仮定することが必要→空間的斉一性(uniformity)

（3）**設計**　これまでに存在していないメカニズムを社会に組み込んだ場合に、起こり得るシナリオを見つけるために、シミュレーションを利用して繰り返し試行錯誤してみる

斉一性（Uniformity）

「どのような場所、歴史上の時間であっても因果関係の構造は変化せず一定である」

モデルの２面性

- **記述的モデル**：現状をよく表現できるスケッチとしてのモデル。現状の詳細な理解を行いたい場合や、将来の動きを予想したい場合に用いられる。一様性の仮定が明示的または暗黙的に導入される。
- **規範的モデル**：社会的システムにおいて、どのような行動様式を有する要素、およびそれらの関係性を構築すると、どのような結果が得られるのかを問う理想的な形を表現するモデル。

課題

社会をシミュレートする場合、常に記述的モデルと規範的モデルとが混在しており、自分の目的が何であるかに注意を払いながら、社会モデルの構築とシミュレーション、シミュレーション結果の分析を総合に行う必要がある

古典的な因果の問題

① 因果関係 (causal relation)
② 因果の推論 (causal inference)
③ 因果の原則 (causal principle)

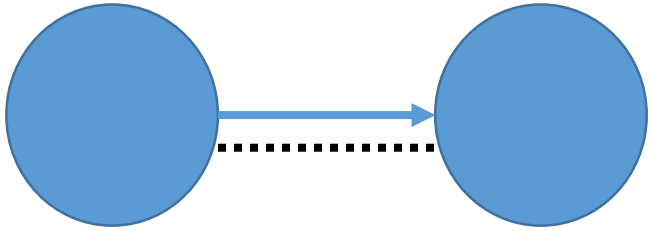
David Hume, A Treatise of Human Nature: Begin an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects, http://www.gutenberg.org/files/4705/4705-h.html

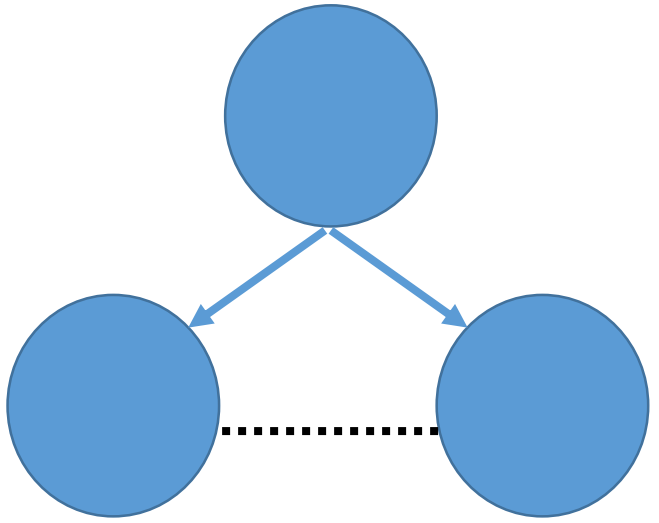
因果関係

- ２つの事象の間で、３つの条件
  - 時間的先行性「原因は結果より時間的に先に起こっていなければならない」
  - 共変関係「２つの事象に強い相関関係が認められるべき」
  - 説明可能性「２つの事象の関係性について因果関係がどうして生じているかについて説明ができる」
- を満たし、かつそれが繰り返し観測される（再現的）ときに我々が認識する2つの事象間の関係性

因果推論の問題点

- 一意性の問題：2つの事象間に因果関係が認められるための3条件を満足しているからといっても、同じ因果関係を生み出す複数の実現可能性が存在している
- 3つめの共通の原因により因果を生じる2つの事象の間には因果が存在しているように見える





帰納の問題(problem of induction)

「帰納的な方法な無限回の観測を経なければ完全に因果関係を説明することができない」

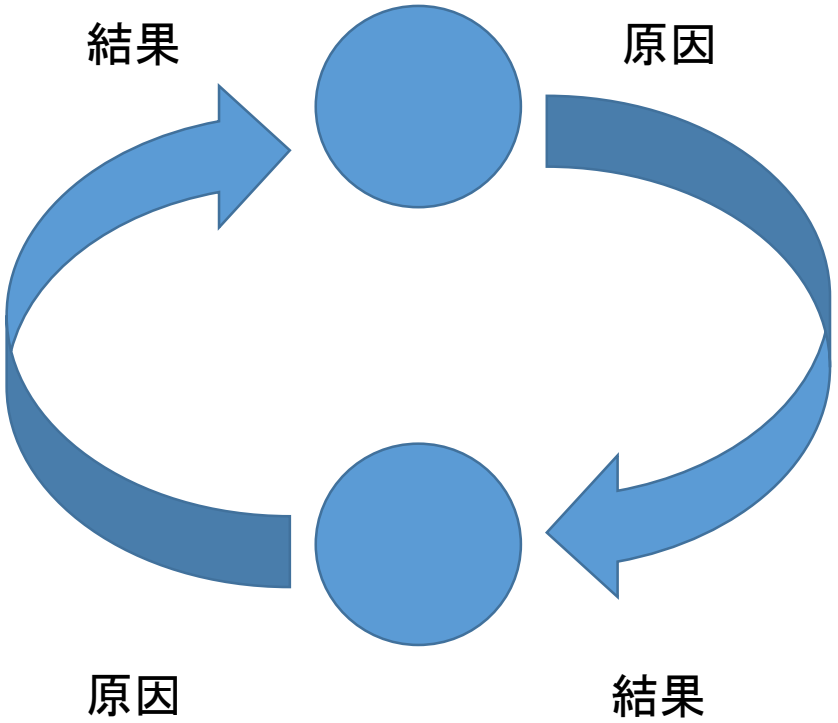
斉一性の原理

- 自然科学においては自然の斉一性原理が仮定できる場合が多いが、社会科学においては社会の斉一性原理は多くの場合成立していない可能性がある

- 強い境界条件を社会に課した場合に、多くの人が一定の原因に対して同じ行為を行う確率が高くなることがある
- 社会シミュレーションを行う場合には、人々の嗜好についてどのような違いがあるかについて、データに基づき事前によく調べておく必要がある

循環的因果の問題

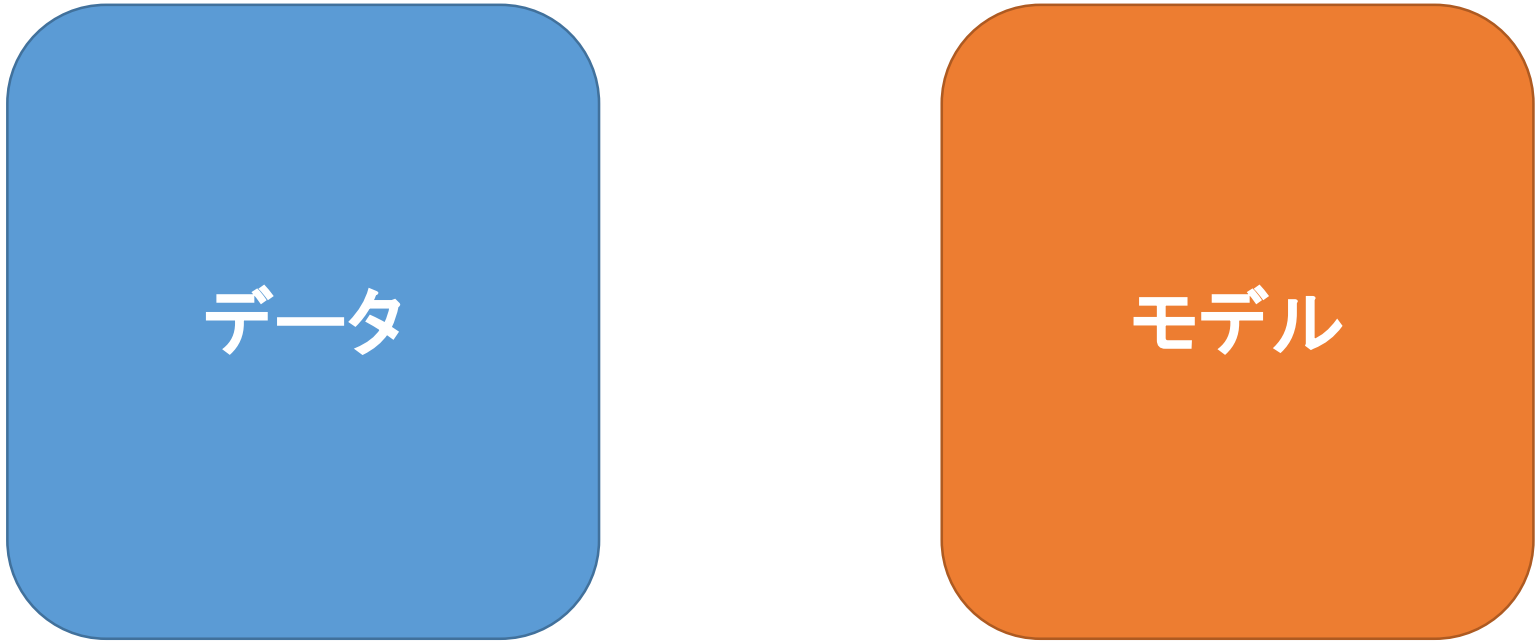
- 因果関係が直線的ではなく環状構造をなしている場合、原因と結果との区別が極めて困難となる



例：都市

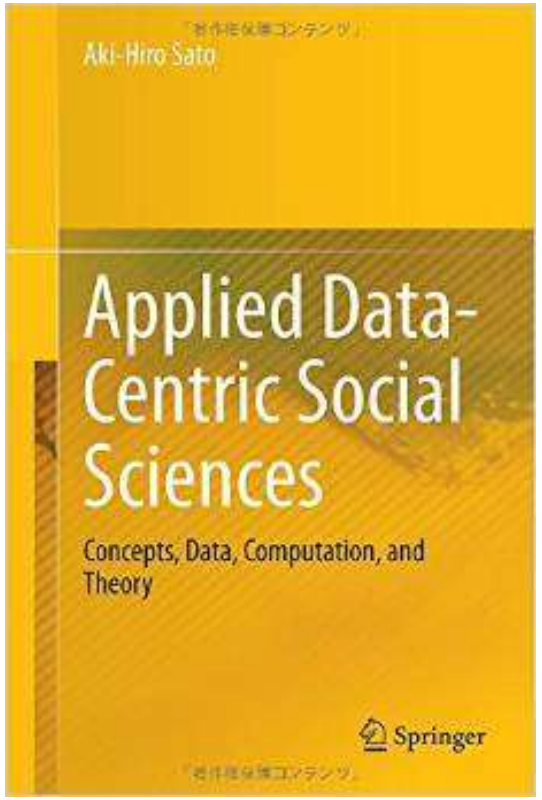
社会シミュレーションにおける課題

- データとモデル双方での積み上げ



Data-Centric Science

Aki-Hiro Sato, “Applied Data-Centric Social Sciences: Concepts, Data, Computation, and Theory,” 2014 (Springer, Tokyo)



I. Introduction

- Introduction
- Framework

II. Methodology

- Mathematical Expressions
- Data in Computers

III. Exemplar Studies

- Risk Assessment of Extreme Events
- Hotel Booking Data
- Tendency in International Air Travels
- Energy Consumption

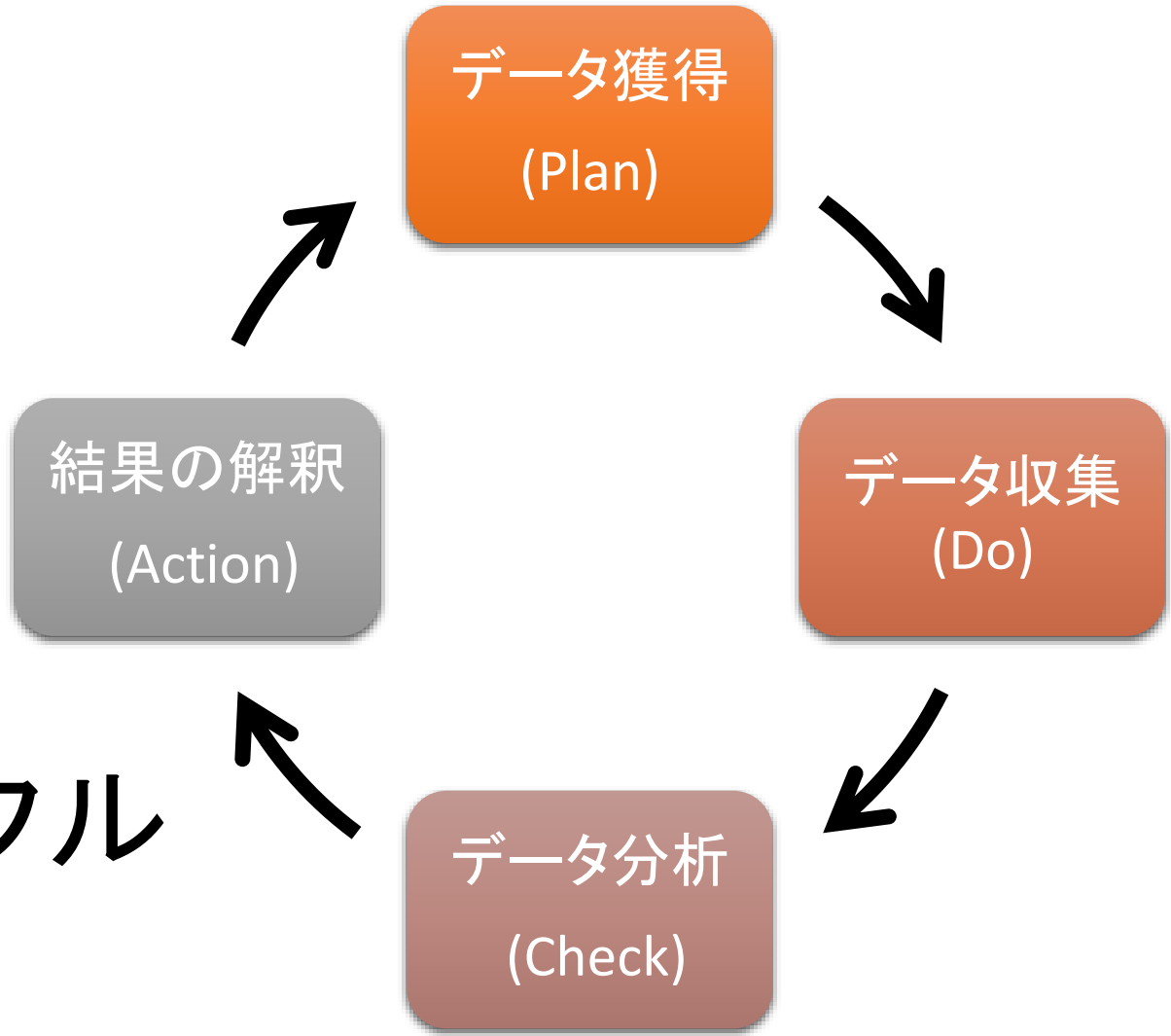
IV. Future Work

- Future Research in Applied Data-Centric Social Sciences

ISBN: 978-4-431-54973-4 (Print) 978-4-431-54974-1 (Online)

データ駆動型研究

- データ獲得
- データ収集
- データ分析
- 結果の解釈



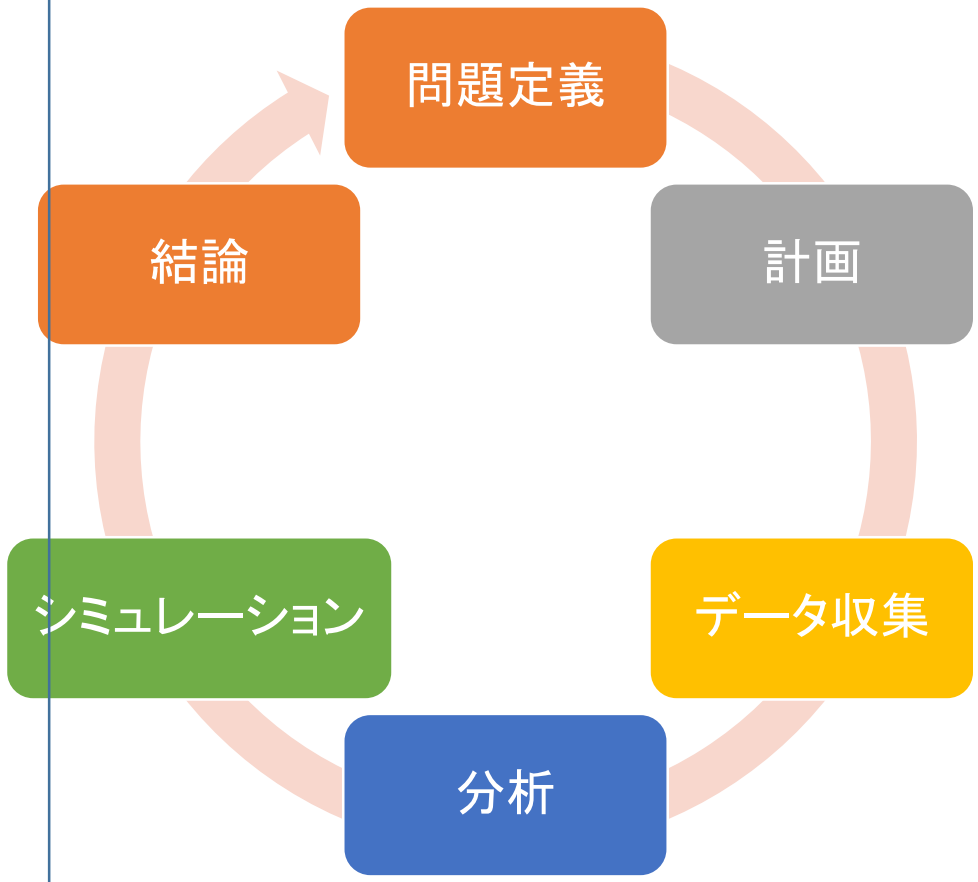
→ CAPDサイクル

PPDASC サイクル

(1) 取り扱いたい現象と関係しそうな小さなデータを用いた研究から始める

(2) PPDASCサイクルを通じてデータ駆動型研究を行う

(3) 得られた結論からデータサイズとデータ種類を増やして再度問題定義をしなおし、データ分析とシミュレーションをビッグデータへと漸近させていく

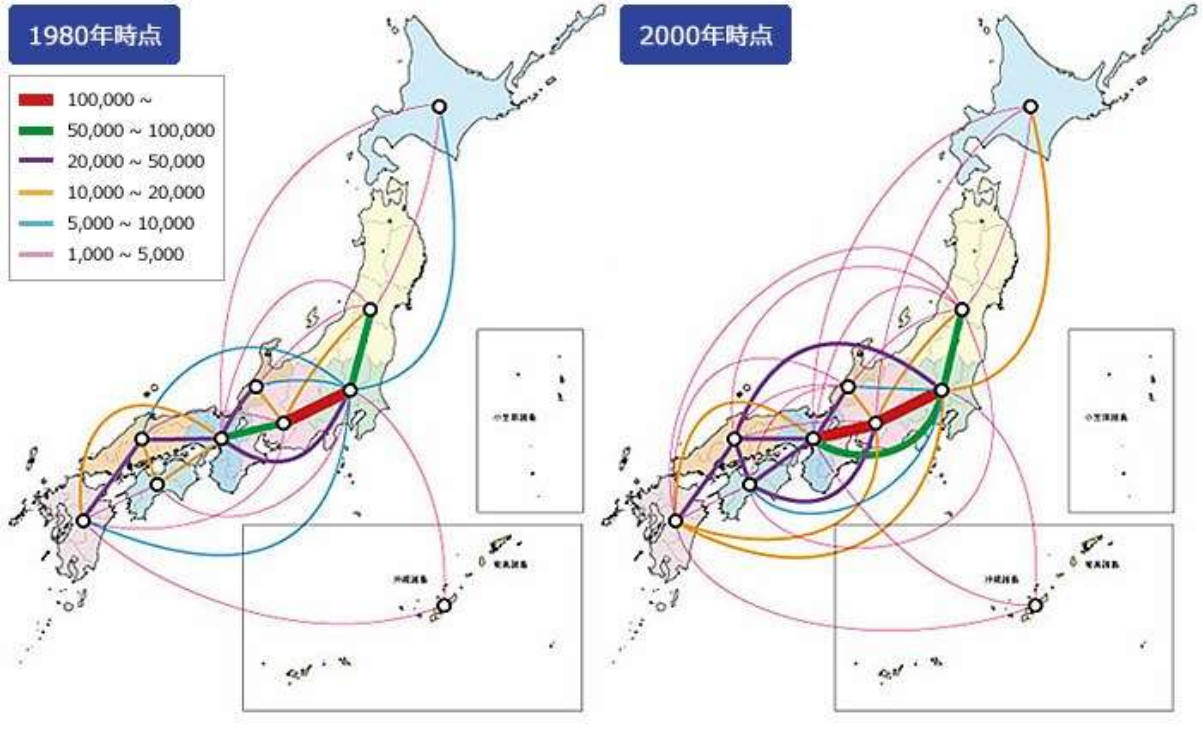


２地点間の人の移動

Aki-Hiro Sato, Hidefumi Sawai, “Relationship between socioeconomic flows and social stocks: case study on Japanese air transportation”, Evolutionary and Institutional Economics Review, Volume 12, Issue 2 (2015) pp 243–263

- 航空機ネットワークにおける２地点間の人の移動量はどのよう傾向にあるのか？
- 移動の活発な地域の経済活動は活発なのか？

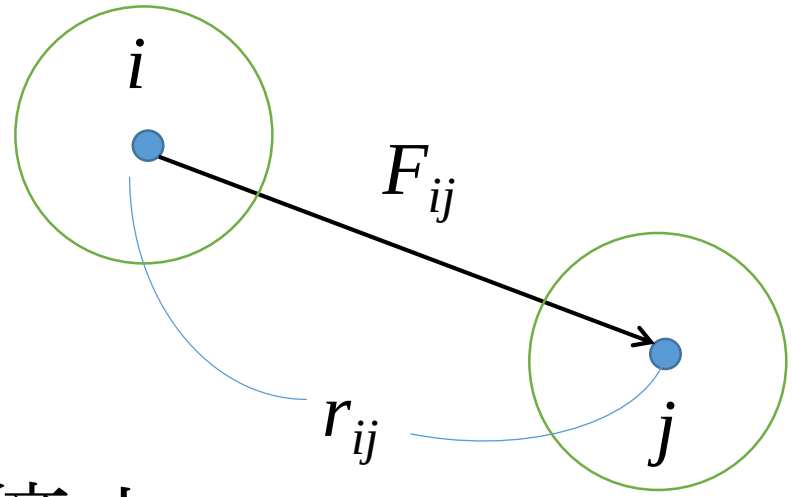
国土交通省国土計画局



重力モデル

Zipf, G.K. The P_1P_2/D hypothesis: on the intercity movement of persons: Am. Sociol. Rev. 11, 677-686 (1946)

- ２地点間($i \rightarrow j$)の移動量 F_{ij}
- 出発地点iの人口$x_i(d)$
- 到着地点jの人口$x_j(d)$
- iとjとの距離r_{ij}



$$F_{ij} = cx_i^{\beta_1}(d)x_j^{\beta_2}(d)r_{ij}^{\beta_3}$$

$x_i(d)$：場所iのd[km]近傍人口

$\beta_0, \beta_1, \beta_2, \beta_3$：回帰係数

重回帰

- 人口と距離の対数に関して線形性を仮定

$$\log F_{ij} = \beta_0 + \beta_1 \log x_i(d) + \beta_2 \log x_j(d) + \beta_3 \log r_{ij}$$

- 最小二乗法

$$E(\beta_0, \beta_1, \beta_2, \beta_3)$$
$$= \sum_{\substack{i,j \\ F_{ij} \neq 0}} \left[\log F_{ij} - \beta_0 - \beta_1 \log x_i(d) - \beta_2 \log x_j(d) - \beta_3 \log r_{ij} \right]^2$$

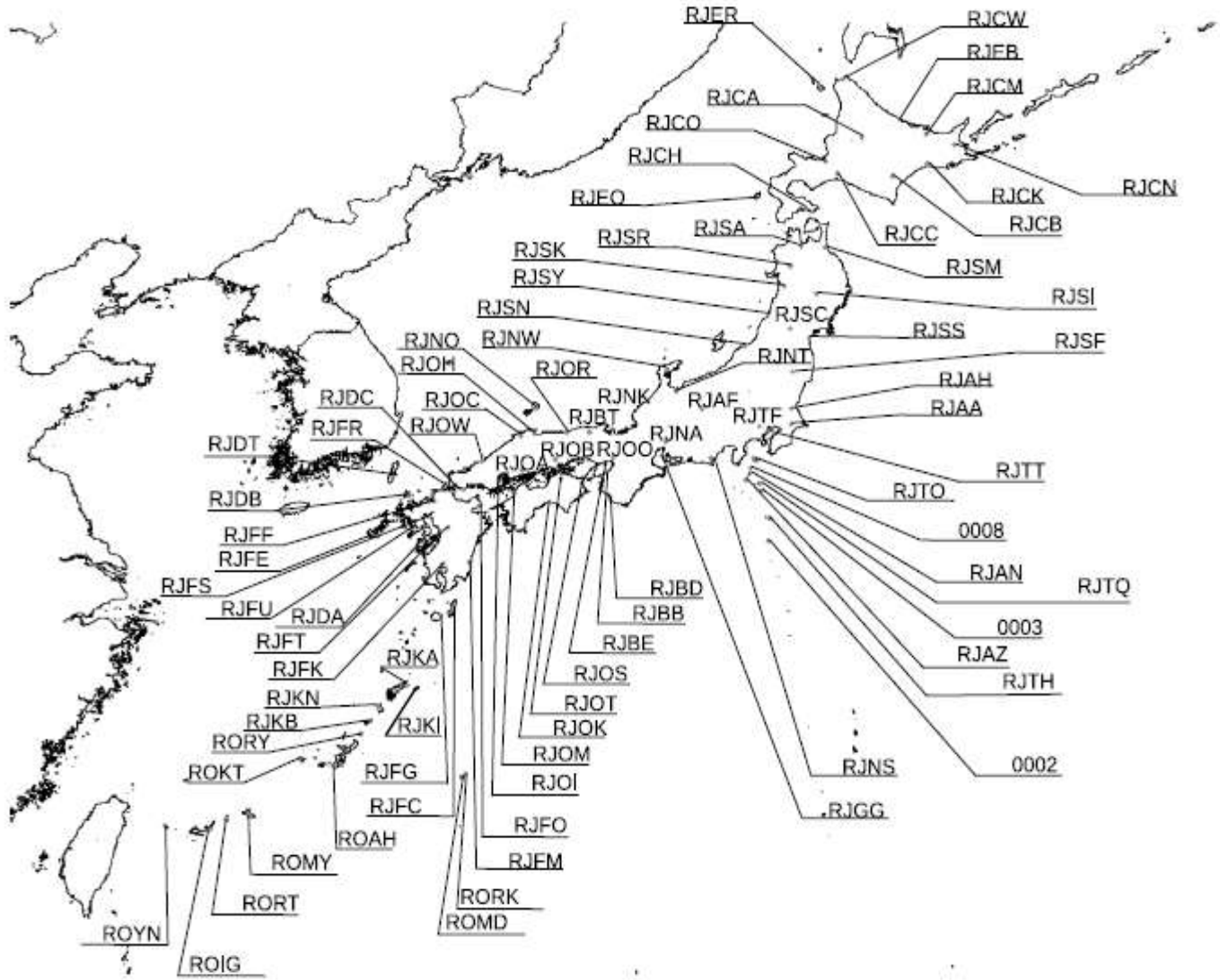
$$(d^*, \hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) = \arg\min_{d, \beta_0, \beta_1, \beta_2, \beta_3} E(d, \beta_0, \beta_1, \beta_2, \beta_3)$$

重力モデルの重回帰係数推計

- 重力モデルを日本の航空輸送統計を用いて検証してみよう

- 重力モデルの係数を計算するために必要となるデータ
  - 人口メッシュデータ
  - ２点間の移動量に関する統計データ
  - 位置データ

日本の空港


RJER
RJCW
RJFB
RJCM
RJCA
RJCO
RJCH
RJCN
RJCK
RJCB
RJEQ
RJSA
RJCC
RJSR
RJSK
RJSM
RJSY
RJSI
RJSN
RJSC
RJSS
RJSF
RJNO
RJNW
RJNT
RJOM
RJOR
RJAH
RJDC
RJOC
RJNK
RJAF
RJAA
RJFR
RJOW
RJBT
RJTF
RJNA
RJDT
RJOB
RJOO
RJOA
RJTT
RJTO
RJDB
0008
RJFF
RJFE
RJAN
RJTQ
RJFS
RJBD
RJFU
RJDA
RJBB
0003
RJFT
RJBE
RJAZ
RJKA
RJFK
RJOS
RJTH
RJKN
RJOT
RJKB
RJKI
RJOK
RORY
RJOM
RJNS
0002
ROKT
RJFG
RJOI
RJFC
RJGG
RJFO
ROAH
RJFM
ROMY
RORK
RORT
ROMD
ROYN
ROIG

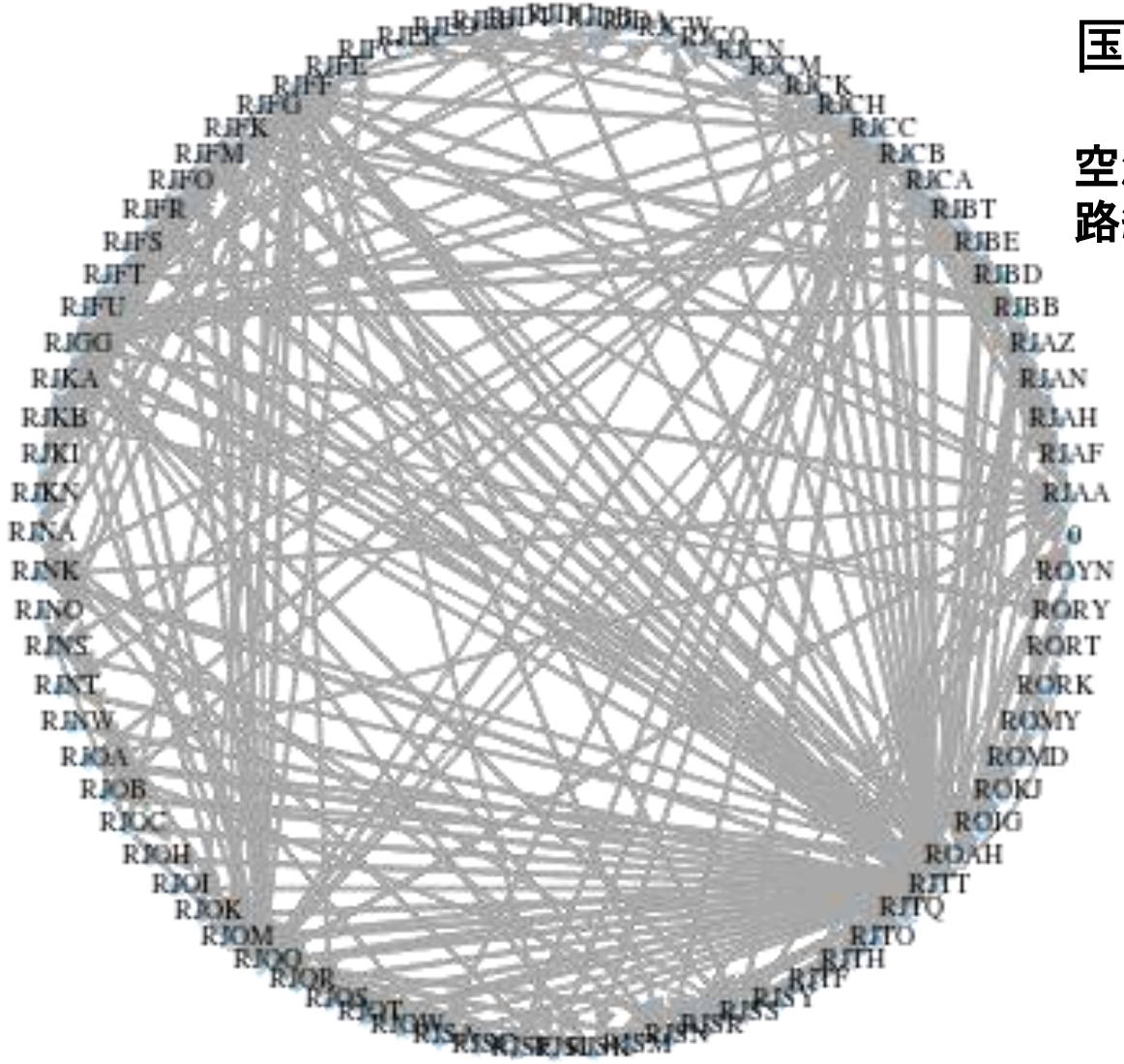
総務省統計局 e-Stat

統計局ポータルサイト e-Stat　http://www.e-stat.go.jp/

公的統計データを利用する

- 2012年航空輸送統計調査（国土交通省）
- 2010年人口（総務省統計局国勢調査）
- 2012年労働者数（総務省統計局経済センサス）
- 2012年事業所数（総務省統計局経済センサス）

日本の航空機ネットワーク（旅客）

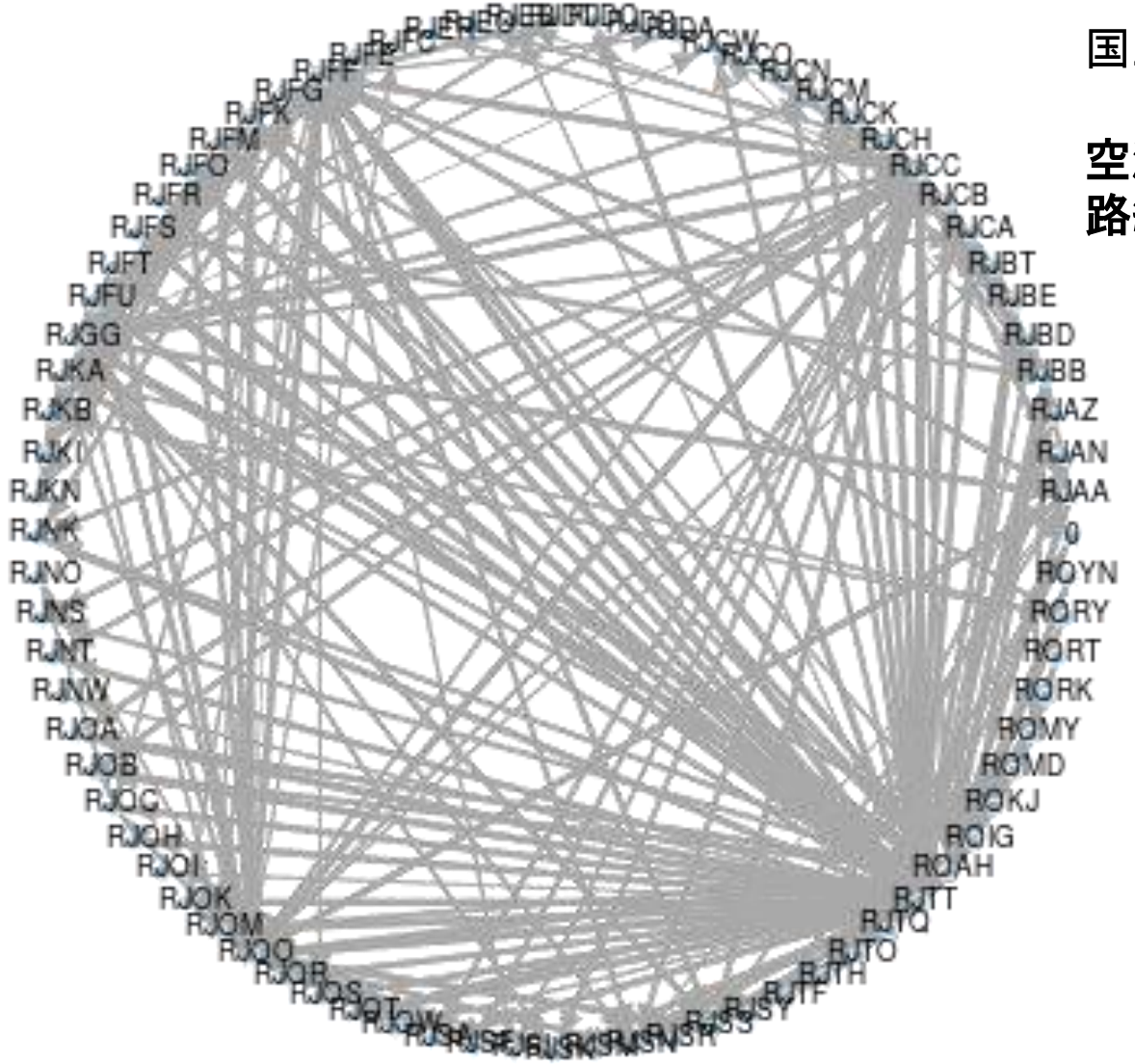


国土交通省2012年航空輸送統計

空港数 86
路線数 476

平均出次数 5.372093
平均入次数 5.372093
平均クラスタ係数 0.469923
大域クラスタ係数 0.205842
入次数エントロピ 2.206731
出次数エントロピ 2.206731
次数アソータティビティ -0.436595
平均次数長 2.473598

日本の航空輸送ネットワーク（貨物）

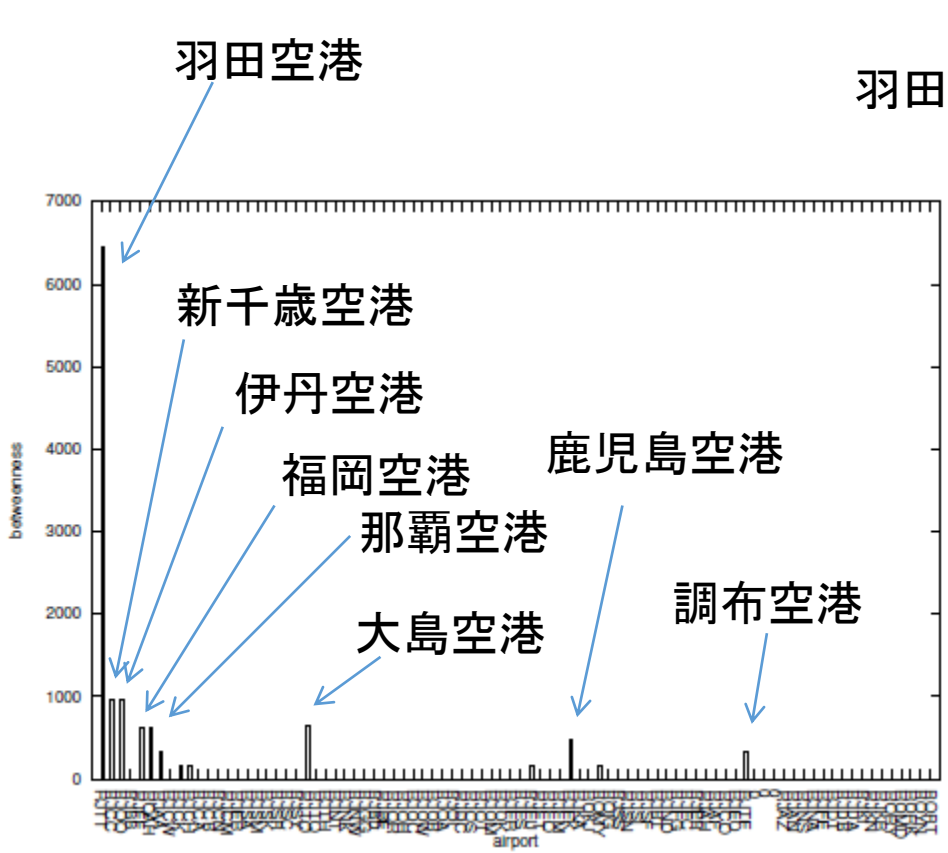


国土交通省2012年航空輸送統計

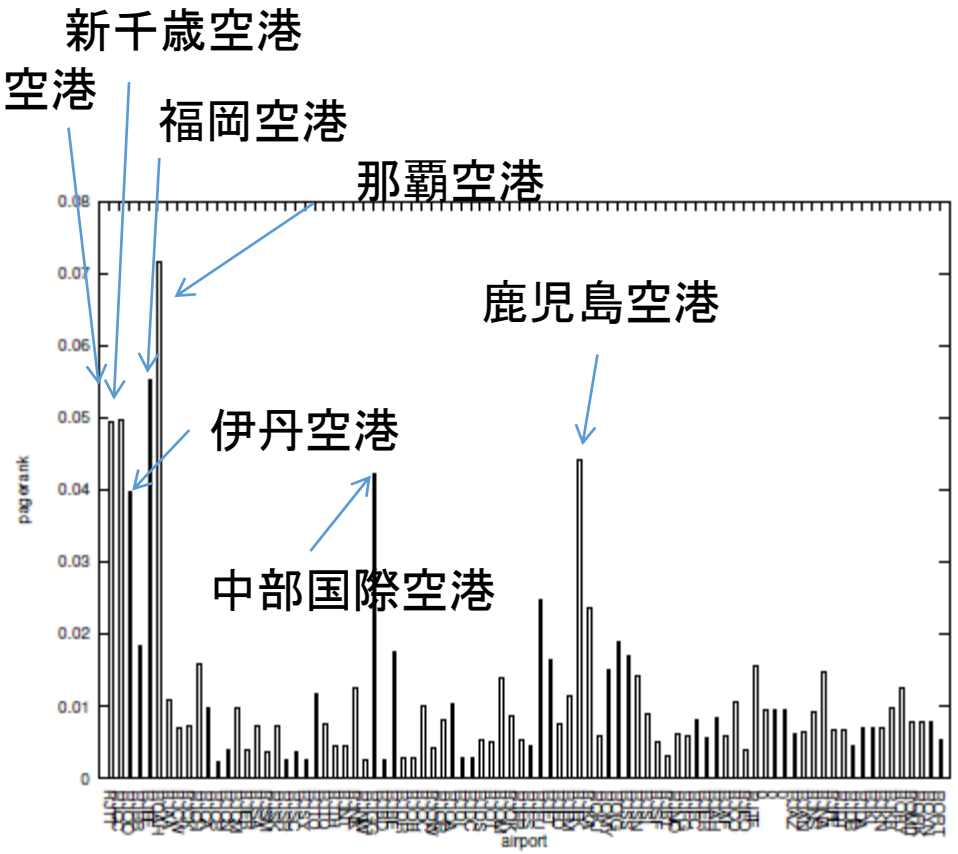
空港数　86
路線数　476

平均出次数 4.804878
平均入次数　4.804878
平均クラスタ係数　0.487389
大域クラスタ係数　0.191510
入次数エントロピ　2.090713
出次数エントロピ　2.090713
次数アソータティビティ　-0.468332
平均次数長 2.494731

日本における推計

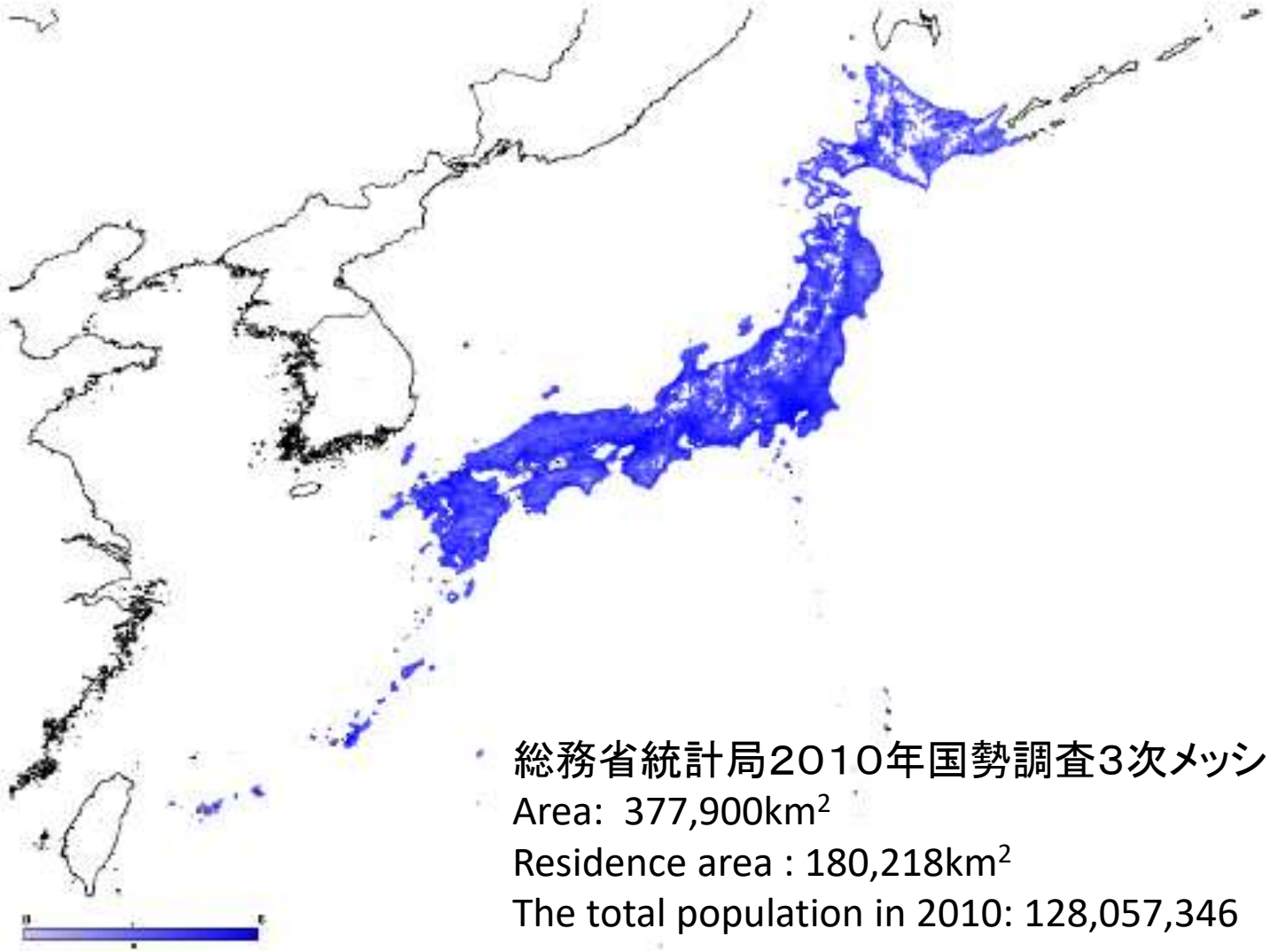


Betweenness centrality



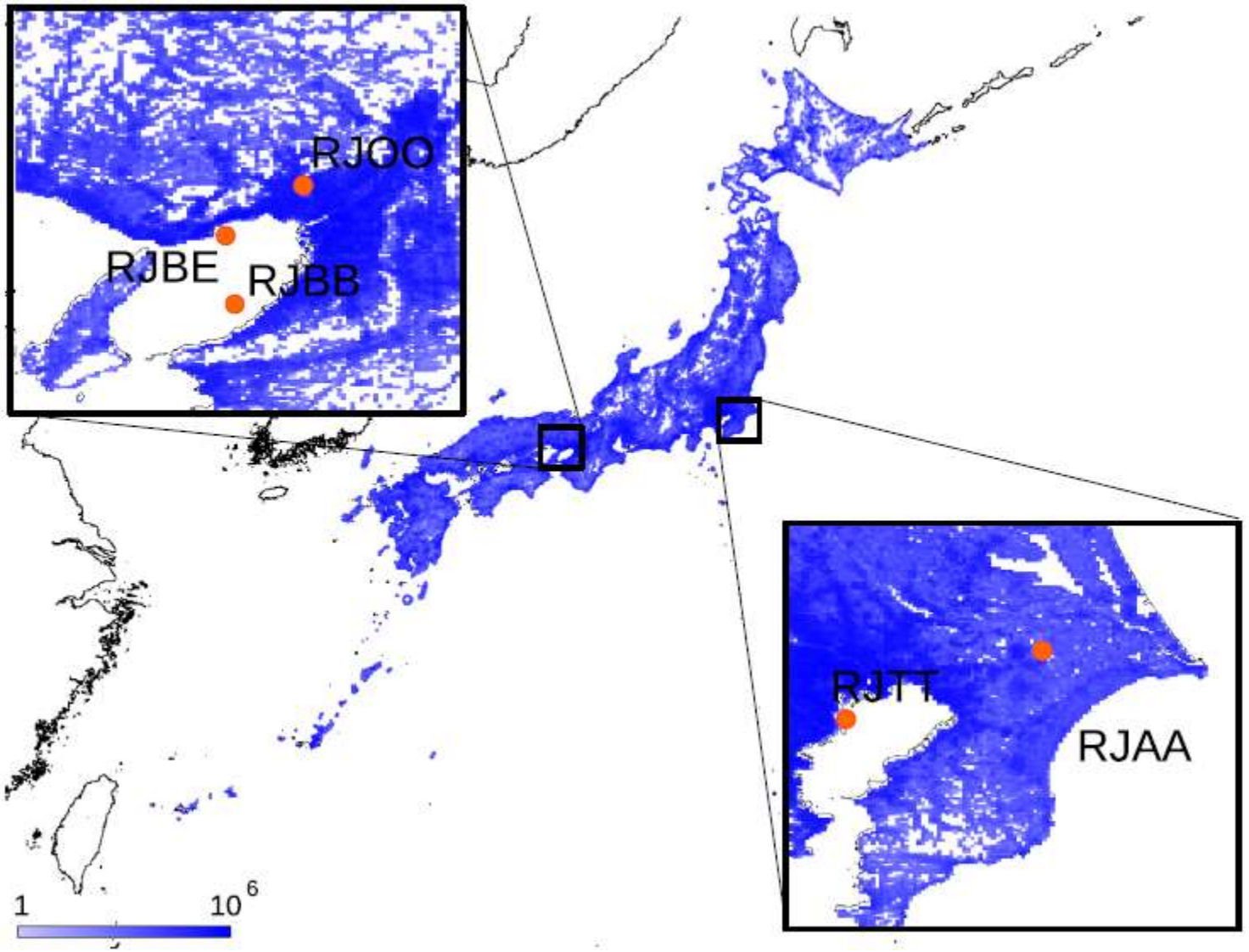
Page Rank

日本の人口



総務省統計局2010年国勢調査3次メッシュデータ
Area: 377,900km^2
Residence area : 180,218km^2
The total population in 2010: 128,057,346

空港周辺の人口



正規方程式

$$\begin{bmatrix} M & \sum \ln x_i(d) & \sum \ln x_j(d) & \sum \ln r_{ij} \\ \sum \ln x_i(d) & \sum (\ln x_i(d))^2 & \sum \ln x_i(d) \ln x_j(d) & \sum \ln x_i(d) \ln r_{ij} \\ \sum \ln x_j(d) & \sum \ln x_i(d) \ln x_j(d) & \sum (\ln x_j(d))^2 & \sum \ln x_j(d) \ln r_{ij} \\ \sum \ln r_{ij} & \sum \ln x_i(d) \ln r_{ij} & \sum \ln x_j(d) \ln r_{ij} & \sum (\ln r_{ij})^2 \end{bmatrix} \begin{bmatrix} \beta_0(d) \\ \beta_1(d) \\ \beta_2(d) \\ \beta_3(d) \end{bmatrix} = \begin{bmatrix} \sum \ln F_{ij} \\ \sum \ln F_{ij} \ln x_i(d) \\ \sum \ln F_{ij} \ln x_j(d) \\ \sum \ln F_{ij} \ln r_{ij} \end{bmatrix}$$

ΣはF_{ij}が0でない全ての空港 i, jについての和を表す
Mは流量データ数

平均二乗誤差

$$MSE(d) = \frac{1}{M-1} \sum_{\substack{i,j \\ F_{ij} \neq 0}} \left(\ln F_{ij} - \beta_0(d) - \beta_1(d) \ln x_i(d) - \beta_2 \ln x_j(d) - \beta(d) \ln r_{ij} \right)^2$$

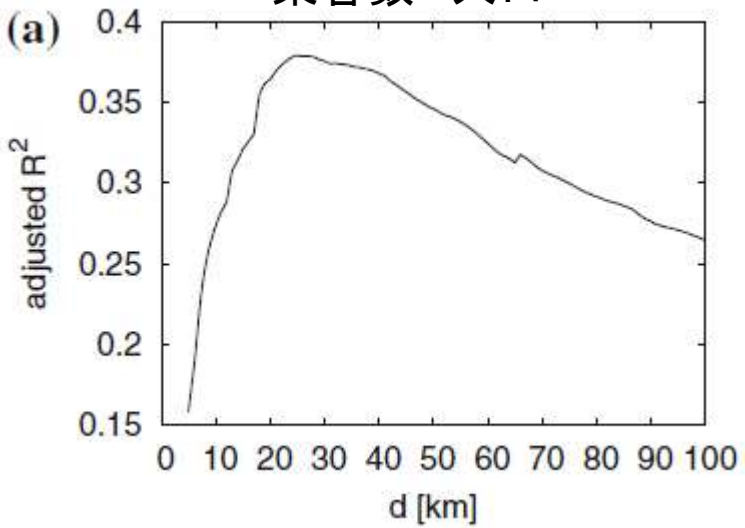
$$d^* = \arg\min_d MSE(d)$$

修正済みR^2

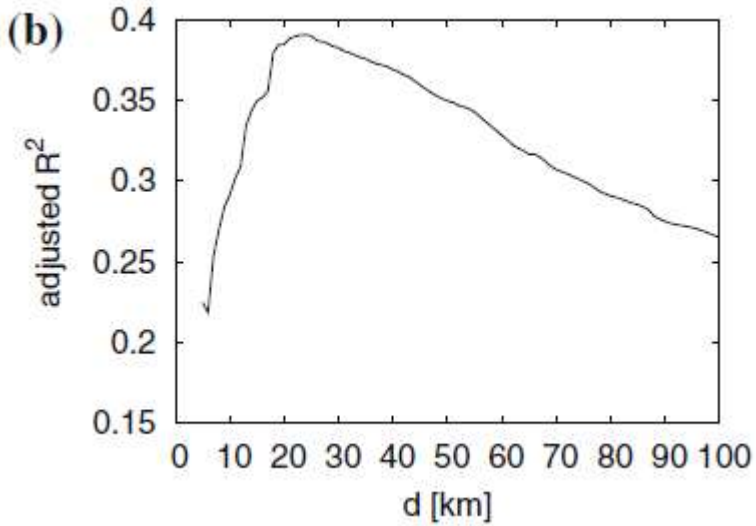
$$\text{adjusted } R^2(d) = 1 - \frac{\sum_{\substack{i,j \\ F_{ij}\neq 0}} \left(\ln F_{ij} - \beta_0(d) - \beta_1(d)\ln x_i(d) - \beta_2 \ln x_j(d) - \beta(d)\ln r_{ij}\right)^2 /(M-5)}{\sum_{\substack{i,j \\ F_{ij}\neq 0}} \left(\ln F_{ij} - \frac{1}{M}\sum_{\substack{i,j \\ F_{ij}\neq 0}} \ln F_{ij}\right)^2 /(M-4)}$$

修正済みR^2

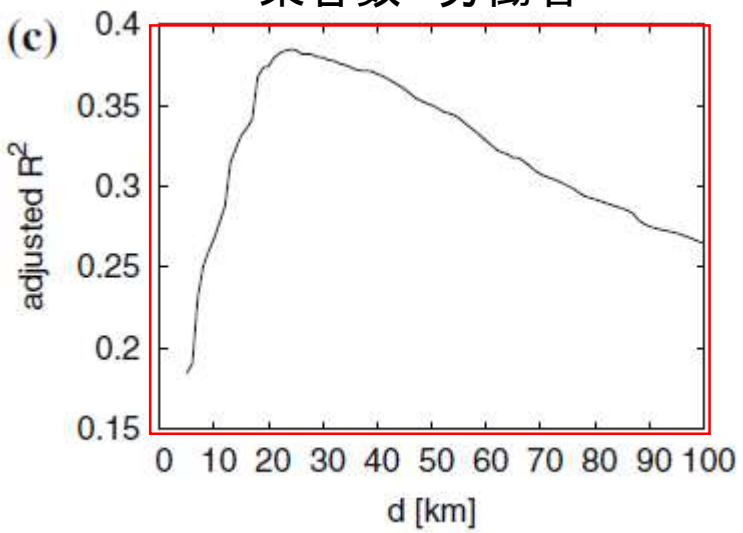
乗客数 ~人口



乗客数 ~事業所



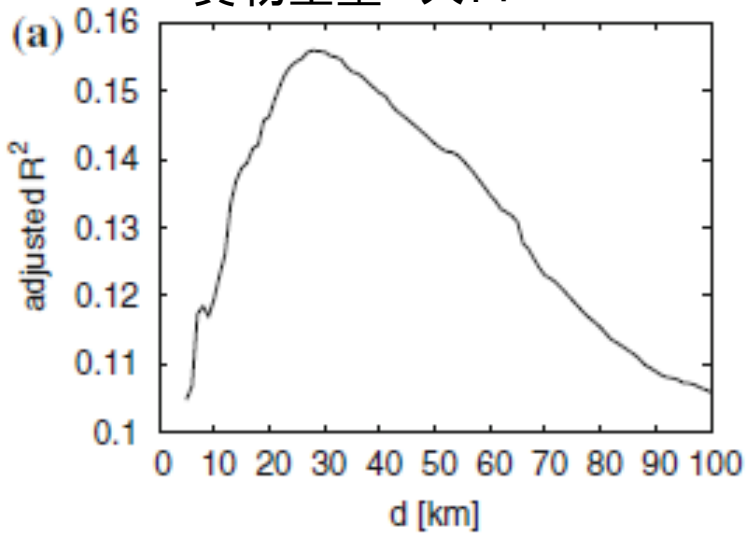
乗客数 ~労働者



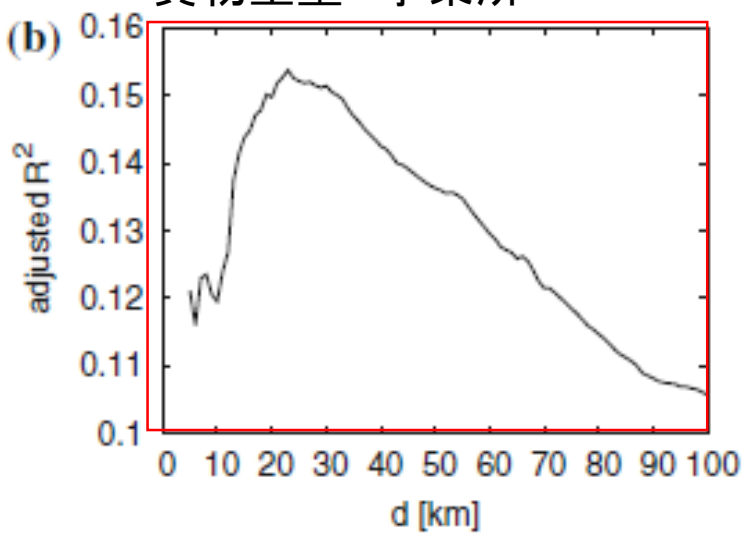
MSE	$\beta_0(d)$	$\beta_1(d)$	$\beta_2(d)$	$\beta_3(d)$
1.314442 (0.389791)	3.001870	0.311012	0.316914	-0.019794
(std.)	0.493901	0.028937	0.028937	0.086524
(t-val)	6.077873	10.74788	10.95185	-0.228768
(Pr[>|t|])	0.000000	0.000000	0.000000	0.819152

修正済みR^2

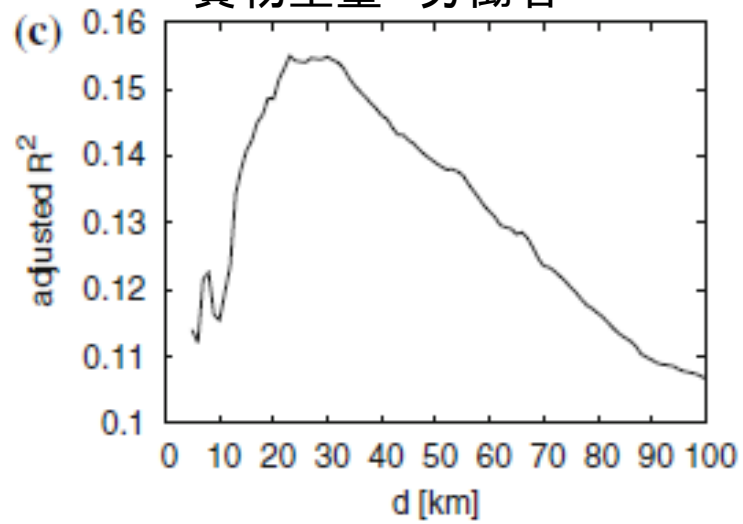
貨物重量 ~人口



貨物重量 ~事業所



貨物重量 ~労働者

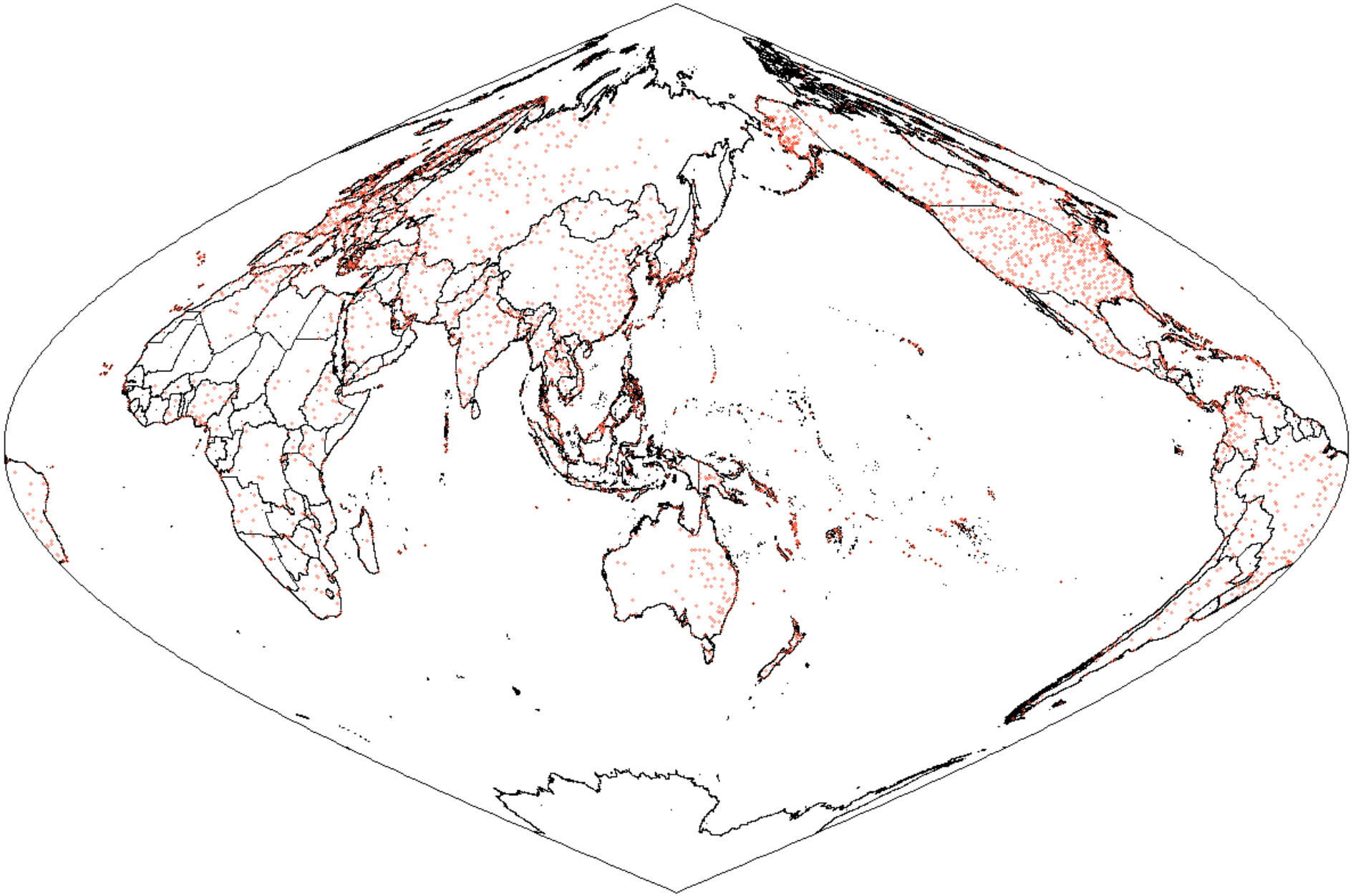


MSE	$\beta_0(d)$	$\beta_1(d)$	$\beta_2(d)$	$\beta_3(d)$
3.159087 (0.154949)	0.964468	0.328385	0.409132	0.410291
(std.)	1.207534	0.082066	0.082241	0.215852
(t-val)	0.798709	4.001457	4.974785	1.900798
(Pr[>|t|])	0.424946	0.000075	0.000001	0.058065

まとめ

- 日本国内における航空輸送について重力モデルのパラメータを推計
- 人流は距離にあまり依存していない
- 物流は距離が遠いほど量が増える
- 適切な空港近傍距離として旅客数に対して25km周辺労働者数、貨物重量に対して23km周辺事業所数を得た
- 空港周辺の25km周辺労働者数と23km周辺事業所数を用いることにより存在しない経路について旅客数と貨物量を予想できる

主要な民間空港



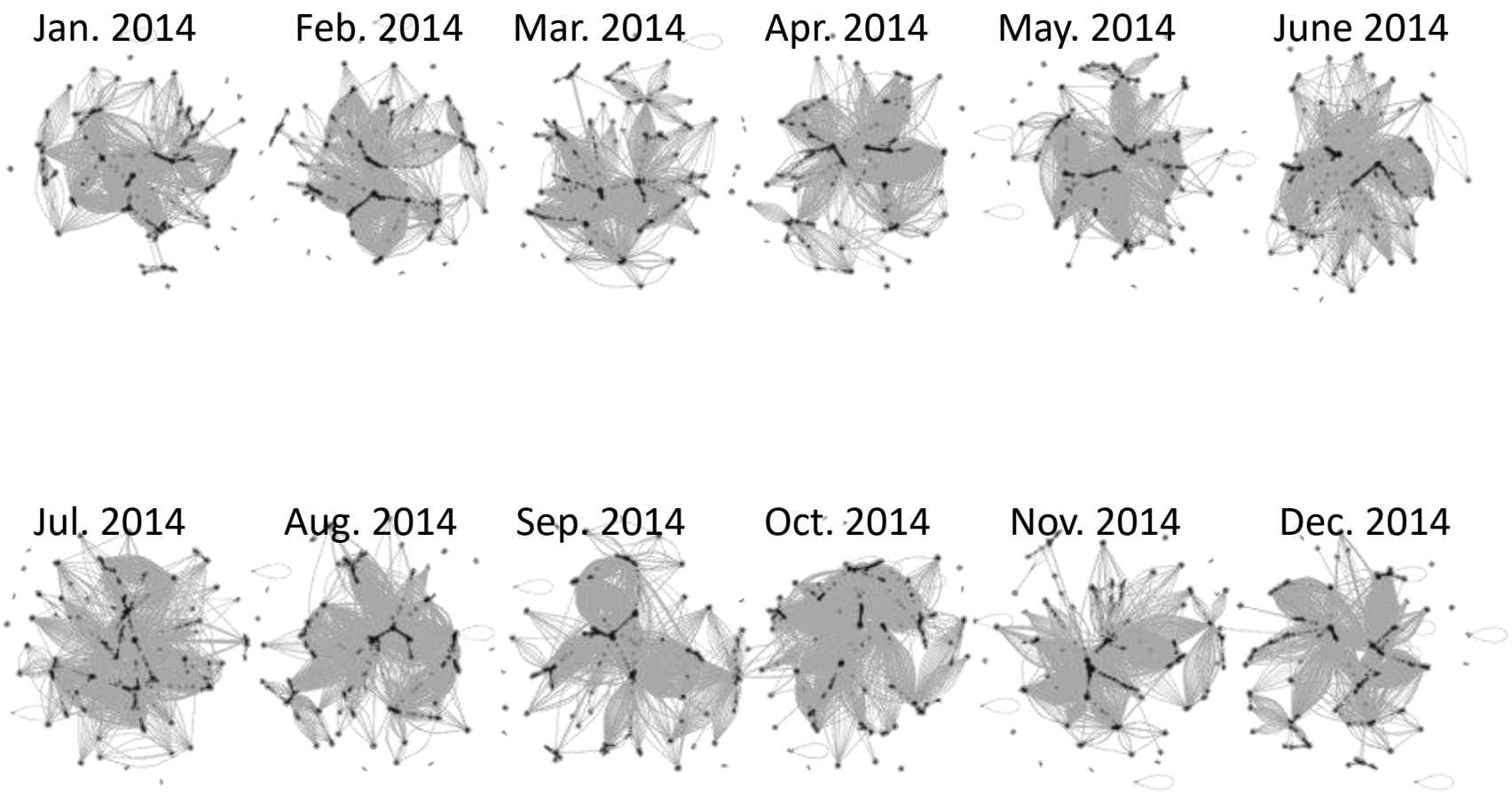
2014年

Aki-Hiro Sato, Hidefumi Sawai , “Risk Assessment for a Global Air Transport System Using Socioeconomic-Technological-Environmental Databases,” 2016 IEEE 40th Annual Computer Software and Applications Conference (2016) pp. 572-581

- 3916空港
- 4,372,945,664座席/年
- 32,159,566便/年
- 54,073経路
- ノード密度 d =54,073/15,331,140=0.3527%

mean in-degree	**13.80722**
mean out-degree	13.80722
average path length	4.031476
assortativity	0.003953
Average betweenness weighted by seat	36044.5 (seats)

世界航空輸送ボリュームデータ

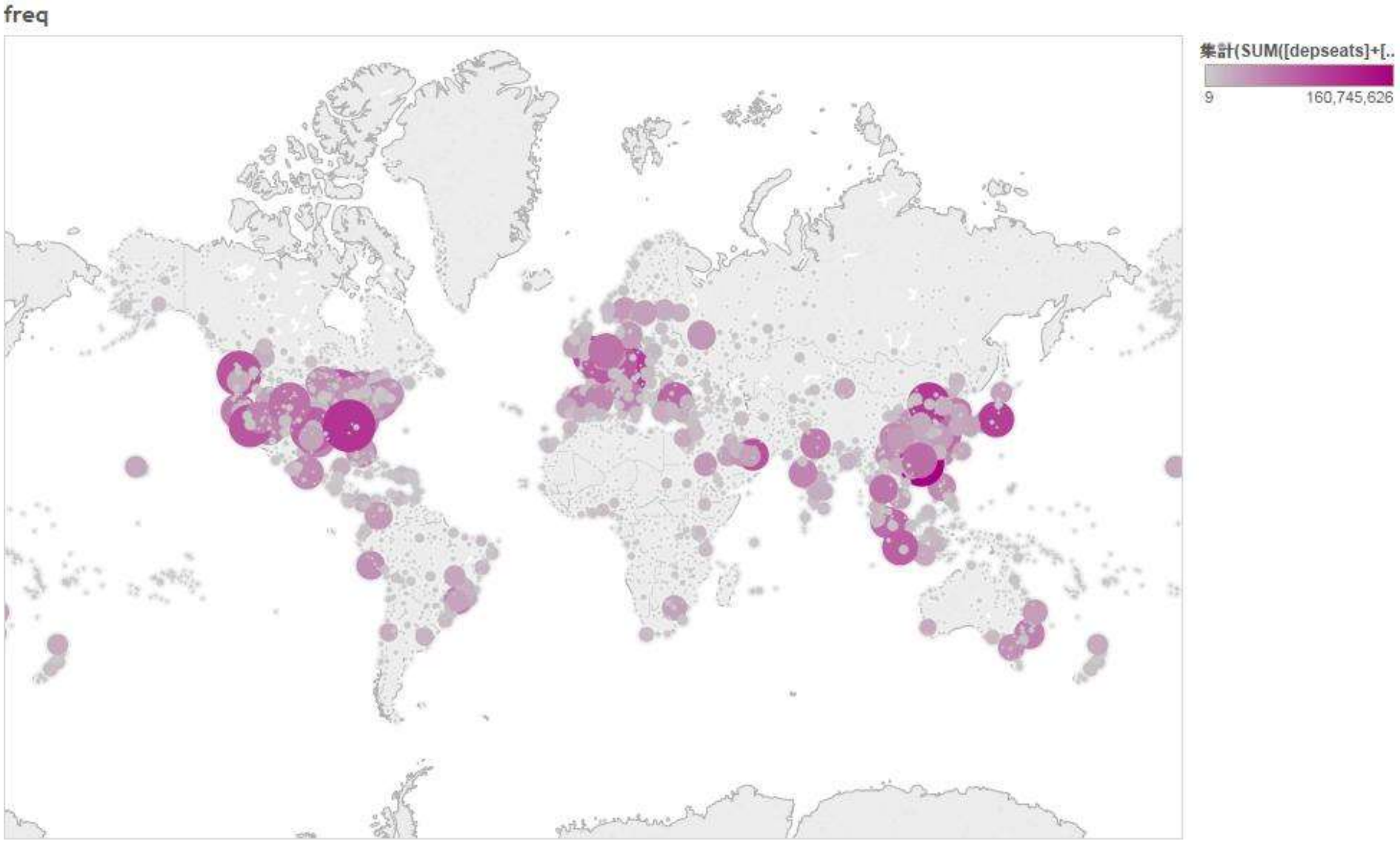


世界航空輸送ボリュームデータ

期間	空港数	主便数	座席総数
Jan. 2014	3,673	117,862	349,054,413
Feb. 2014	3,664	112,622	318,905,914
Mar. 2014	3,683	131,380	357,561,105
Apr. 2014	3,685	113,021	359,074,363
May. 2014	3,670	120,818	374,199,404
Jun. 2014	3,690	130,050	374,011,449
Jul. 2014	3,687	124,173	397,190,537
Aug. 2014	3,690	129,678	398,874,253
Spt. 2014	3,696	124,844	375,040,614
Oct. 2014	3,671	141,577	380,620,602
Nov. 2014	3,661	116,156	350,869,864
Dec. 2014	3,640	115,704	365,761,938
total	3,670	1,477,885	4,401,164,456

発着数と座席数

色の濃さ：座席
〇の大きさ：フライト数

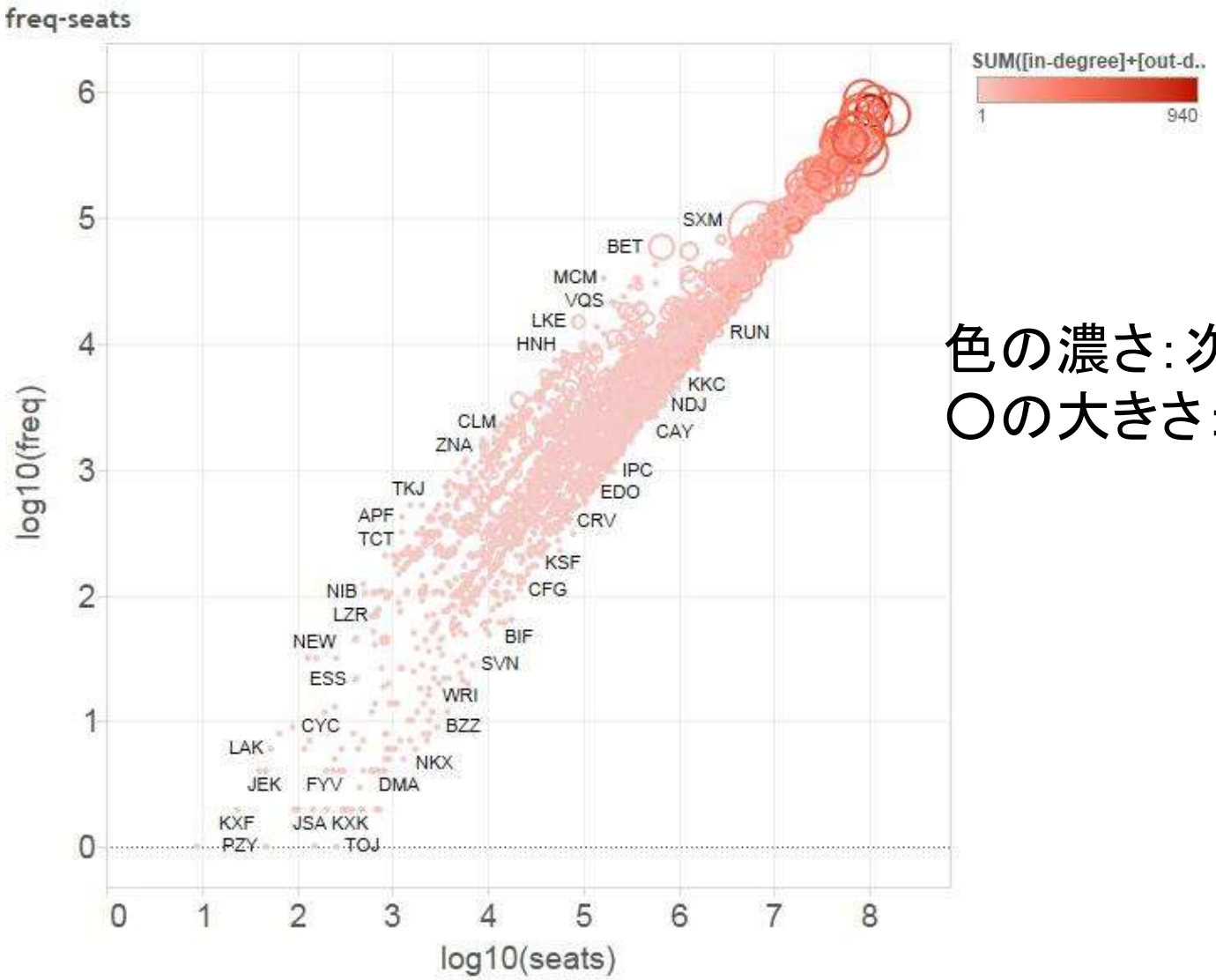


発着数と座席数（上位10位）

IATA	発着数
ORD	879,754
ATL	850,532
MUC	707,362
DFW	661,066
HKG	656,488
SEA	647,878
LAX	585,608
PEK	568,564
DEN	551,672
CLT	513,991

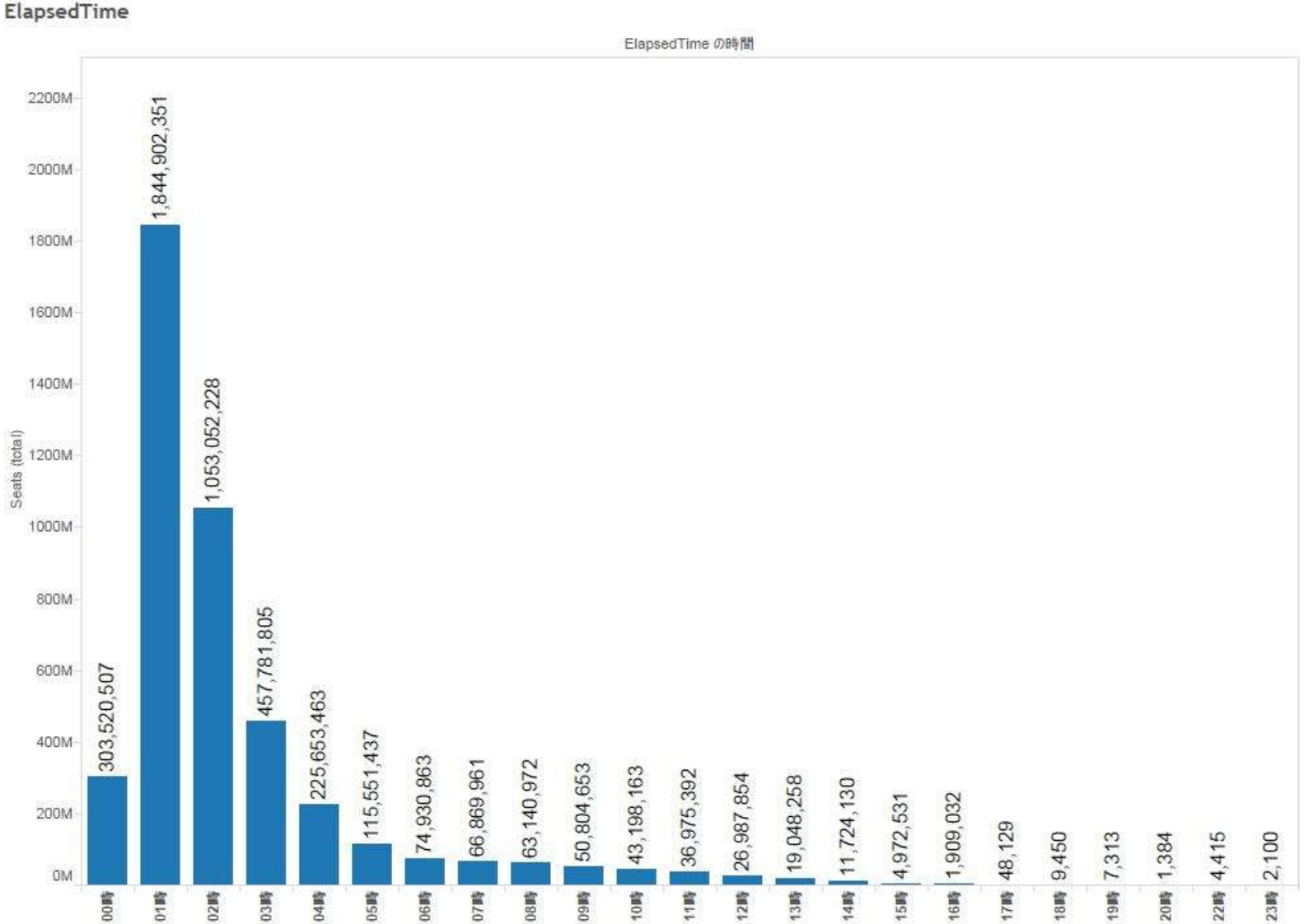
IATA	座席数
HKG	160,745,626
ATL	112,435,142
PEK	109,252,996
MUC	107,249,376
HND	102,231,892
LHR	93,816,439
DXB	89,882,181
SEA	85,930,802
ORD	85,304,177
LAX	84,473,390

座席数とフライト数との関係

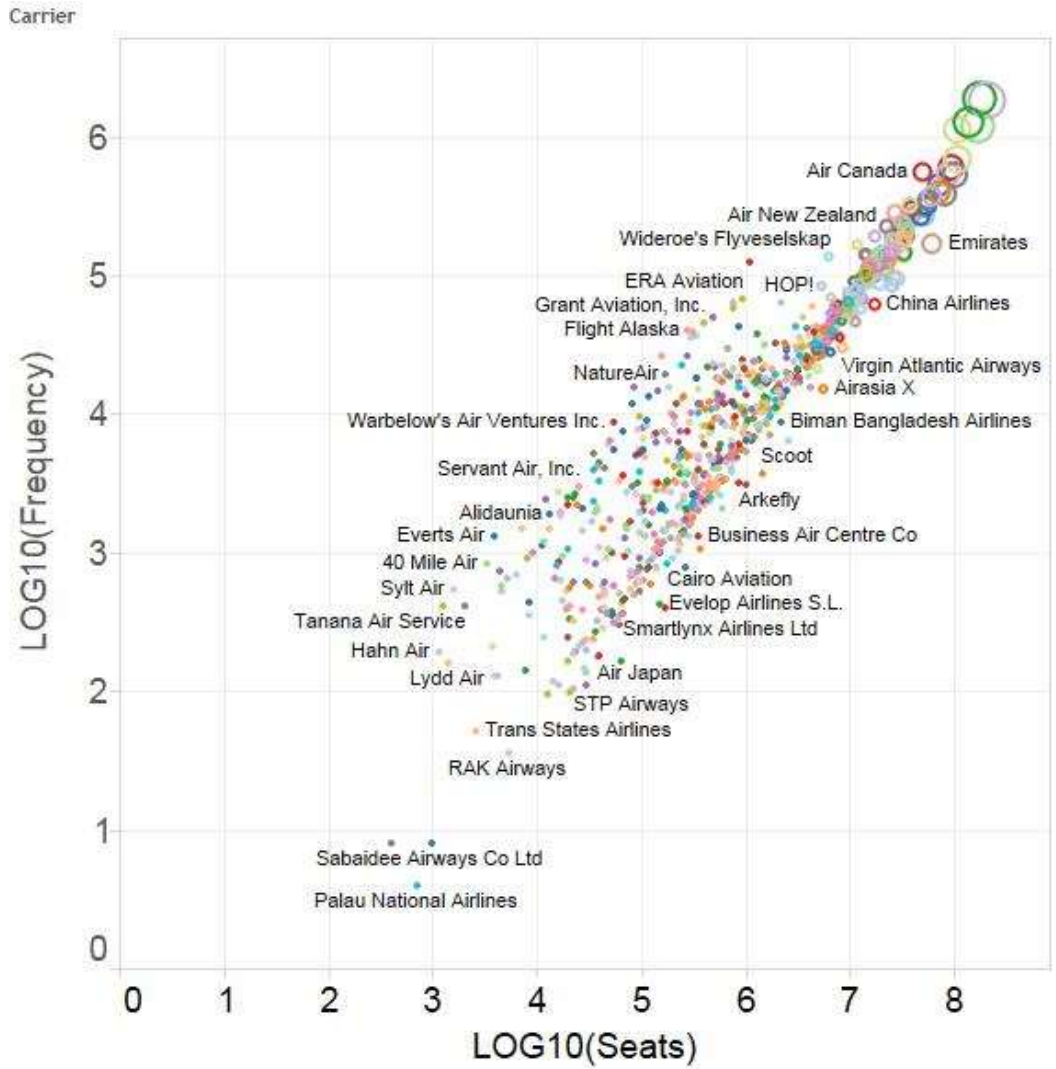


色の濃さ：次数
〇の大きさ：媒介中心性

飛行時間と座席数


ElapsedTime
ElapsedTime の時間
Seats (total)
2200M
2000M
1800M
1600M
1400M
1200M
1000M
800M
600M
400M
200M
0M
303,520,507
1,844,902,351
1,053,052,228
457,781,805
225,653,463
115,551,437
74,930,863
66,869,961
63,140,972
50,804,653
43,198,163
36,975,392
26,987,854
19,048,258
11,724,130
4,972,531
1,909,032
48,129
9,450
7,313
1,384
4,415
2,100
00時
01時
02時
03時
04時
05時
06時
07時
08時
09時
10時
11時
12時
13時
14時
15時
16時
17時
18時
19時
20時
22時
23時

航空会社ごとの座席数とフライト数



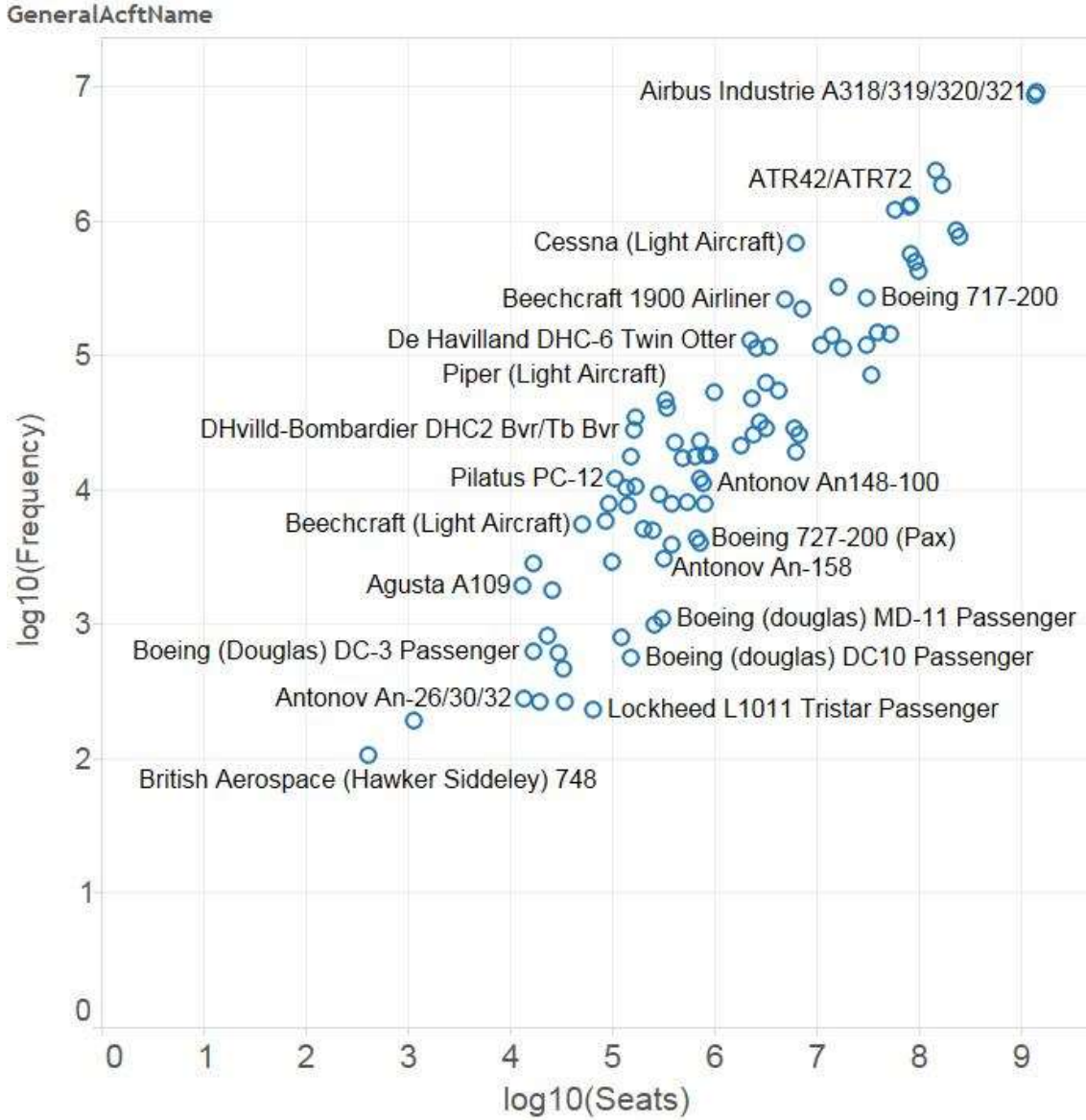
色：航空会社
〇の大きさ：フライト数

航空会社上位10位

航空会社	フライト数
United Airlines	1,868,789
Delta Air Lines	1,846,443
American Airlines	1,269,321
Southwest Airlines	1,179,597
US Airways	1,118,763
China Southern Airlines	679,864
China Eastern Airlines	589,804
Air Canada	562,856
Lufthansa German Airlines	540,389
Ryanair	531,539

航空会社	座席数
Delta Air Lines	209,664,779
United Airlines	174,542,283
Southwest Airlines	170,533,380
American Airlines	139,072,254
US Airways	109,935,355
China Southern Airlines	107,837,110
Ryanair	100,385,211
China Eastern Airlines	94,176,702
Lufthansa German Airlines	86,750,566
All Nippon Airways	79,643,447

機種ごとの座席数とフライト数

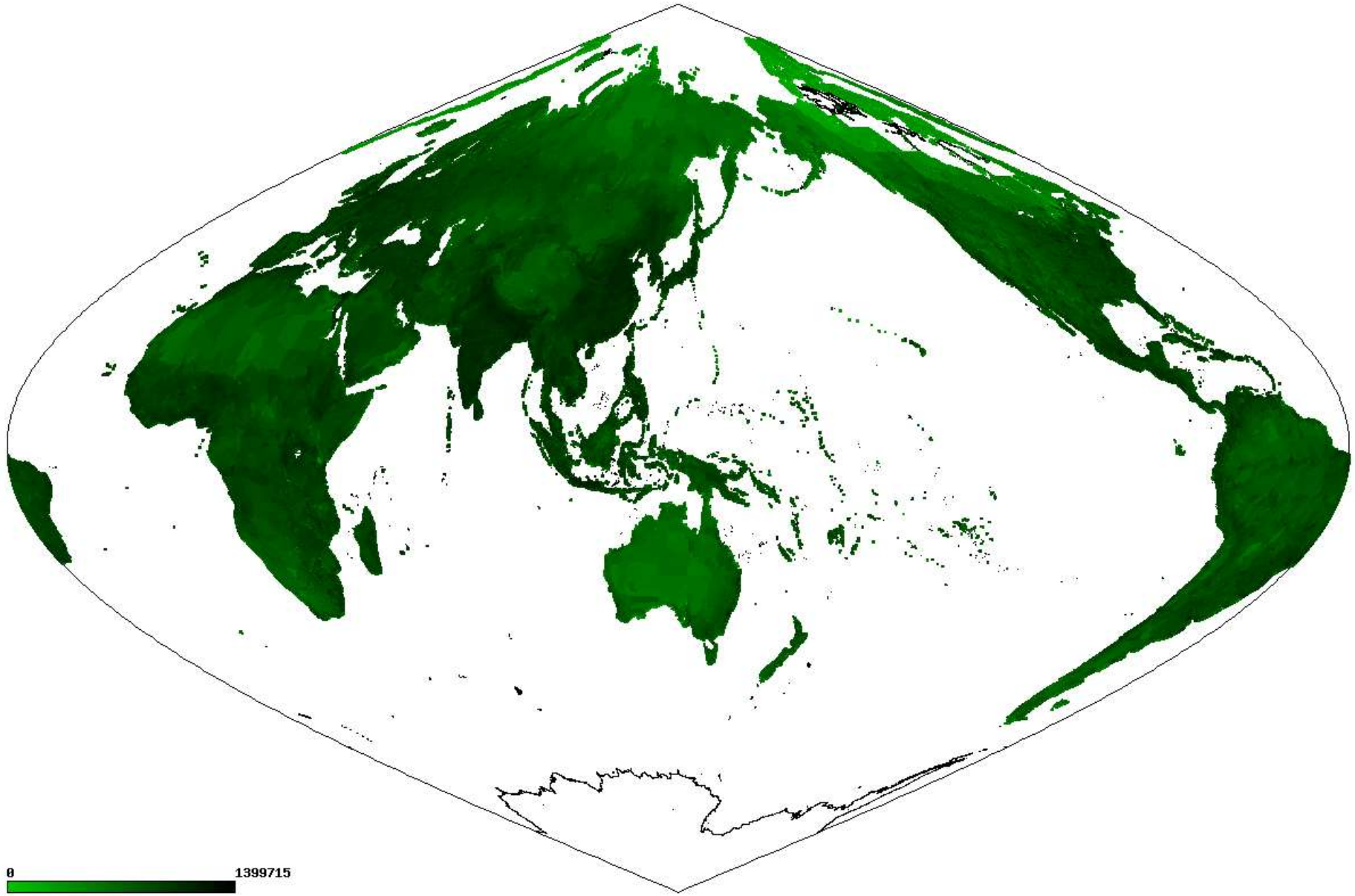


座席数とフライト数上位10位

GeneralAcftName	Frequency	Seats (total)
Airbus Industrie A318/319/320/321	9,041,065	1,431,208,910
Boeing 737 Passenger	8,536,235	1,348,944,037
Canadair Regional Jet	2,343,452	146,873,444
Embraer 170/195	1,849,159	171,507,170
ATR42/ATR72	1,309,327	82,479,339
DHvilld-Bombardier DHC8 Dsh 8	1,275,720	79,578,608
Embraer RJ 135/140/145	1,193,359	58,088,426
Airbus Industrie A330	831,138	229,760,646
Boeing 777 Passenger	762,291	248,147,651
Cessna (Light Aircraft)	670,483	6,243,622

GeneralAcftName	Frequency	Seats (total)
Airbus Industrie A318/319/320/321	9,041,065	1,431,208,910
Boeing 737 Passenger	8,536,235	1,348,944,037
Boeing 777 Passenger	762,291	248,147,651
Airbus Industrie A330	831,138	229,760,646
Embraer 170/195	1,849,159	171,507,170
Canadair Regional Jet	2,343,452	146,873,444
Boeing 767 Passenger	423,416	100,328,575
Boeing 757 (Passenger)	485,573	91,700,175
Boeing (douglas) MD-80	565,452	83,160,396
ATR42/ATR72	1,309,327	82,479,339

世界の2.5分人口メッシュデータ



回帰分析

$$\ln S_{ij} = \beta_0 + \beta_1 \ln x_i(d) + \beta_2 \ln x_j(d) + \beta_3 \ln D_{ij}$$

S_{ij}: 空港iから空港jの輸送座席数
$x_i(d)$: 空港iの周辺d [km]における人口
D_{ij}: 空港iと空港jとの大圏距離 [km]
$\beta_0, \beta_1, \beta_2, \beta_3$: 回帰係数

推定値

2014年1年間の全世界の航空輸送データ (OAG Inc.)

空港数	経路数	フライト総数	座席総数	ノード密度	平均経路長	平均次数
3,916	54,073	32,159,566	4,372,945,664	0.3527%	4.03	13.80722(入次数） 13.80722(出次数)

d [km]	Adj. R2	β_0	β_1	β_2	β_3
36	0.163838	3.861518	0.254348	0.255408	-0.147192

まとめ

- データの獲得収集、整理、管理（データの保守と管理）
- 巨大データベースと並列計算技術の連携
- データと計算プログラム双方のデバッグ

- データ検証と信頼性の高いソフトウェア開発の双方が重要（誤ったデータをどのように正しく計算しても誤った結果しか得られない）
- 巨大データベースと並列計算技術の融合が必要