

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

キャノングローバル戦略研究所
京都大学大学院情報学研究科

2017年9月25日～26日 「経済・社会への分野横断的研究会」

もくじ

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1章 「持続可能な社会」をシミュレーションする——「共有地の悲劇」をめぐる規範と信頼（遠藤 薫）
- 第2章 エージェント・ベース・モデリングの楽しさと難しさ（寺野 隆雄）
- 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）
- 第4章 ソーシャルメディアにおける情報拡散——どのようにしてデマ情報は蔓延し、収束するのか（栗原 聡）
- 第5章 人工社会が予測する都市の動態（倉橋 節也）
- 第6章 シミュレーション技術を応用した3次元文化財の透視可視化（田中 覚）

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

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、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である。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。モデラーが考える因果関係が観測できていない場所と過去・将来においても成立しているという仮定を必要とする
- **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。マートンのロールモデルが有名。

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)

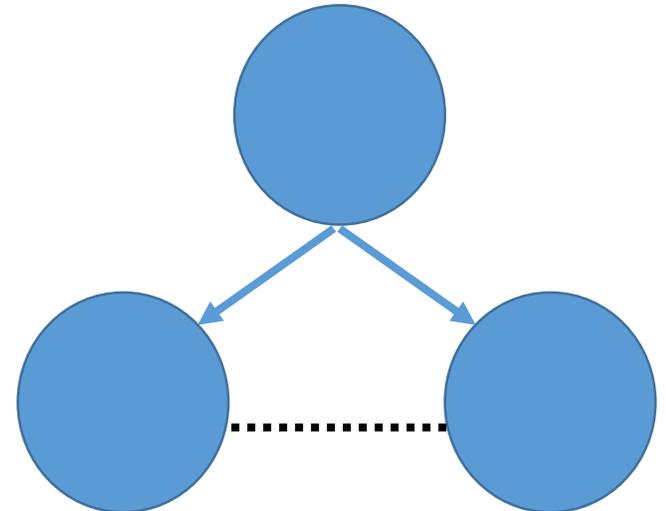
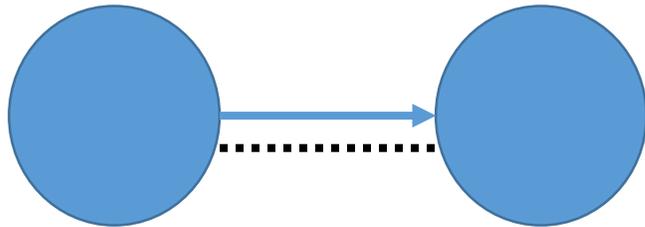
David Hume, A Treatise of Human Nature: Begin an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects, <http://www.gutenberg.org/files/4705/4705-h.html>

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いといけない」
 - 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)

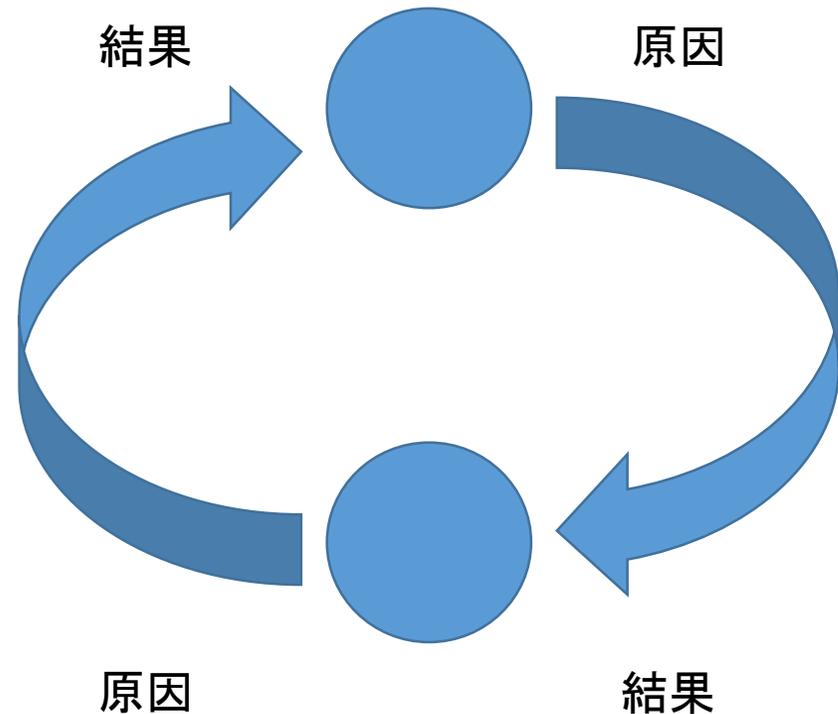
「帰納的な方法な無限回の観測を経なければ完全に因果関係を説明することができない」

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、多くの人々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



データ

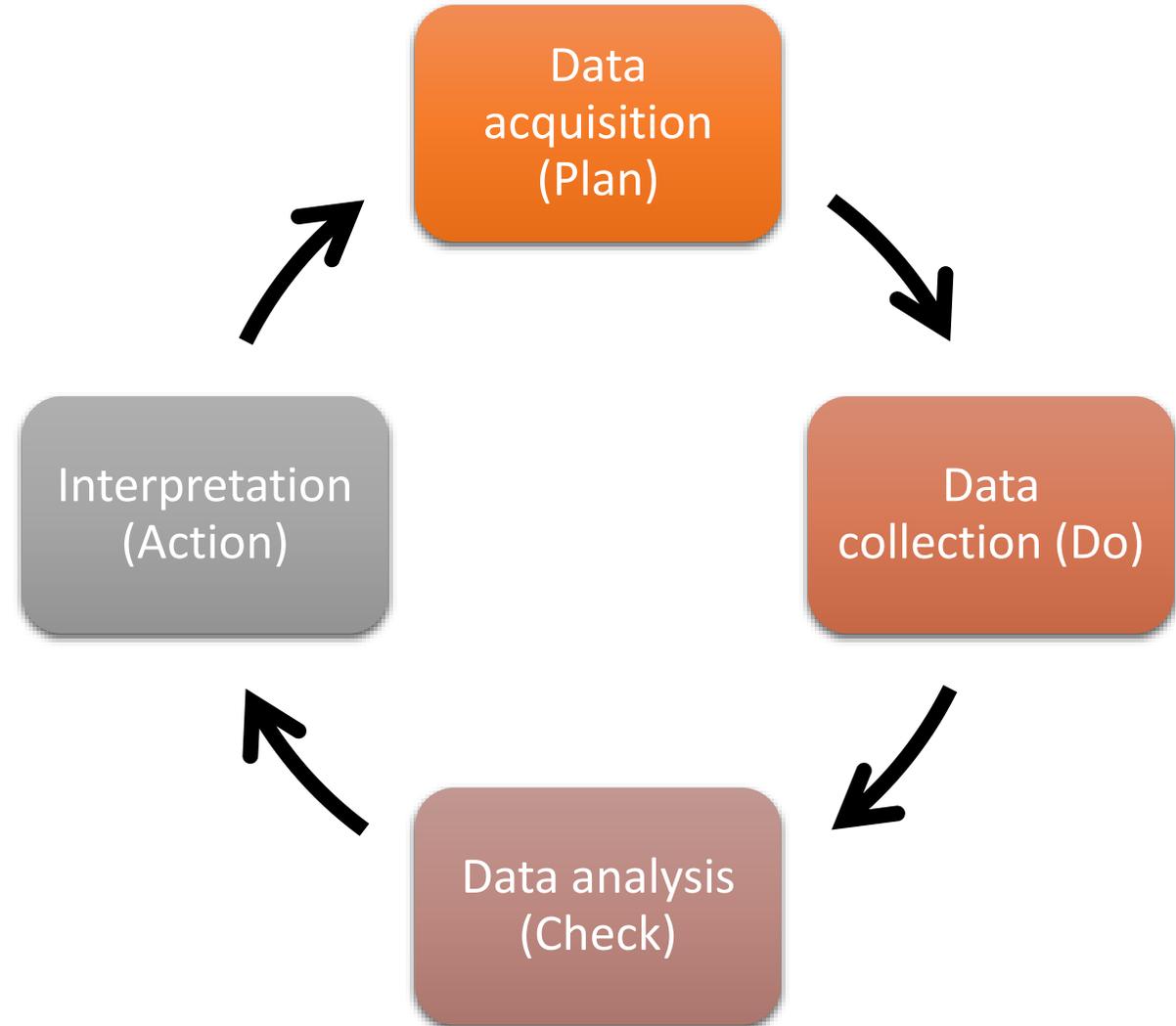


モデル

Data-driven investigation

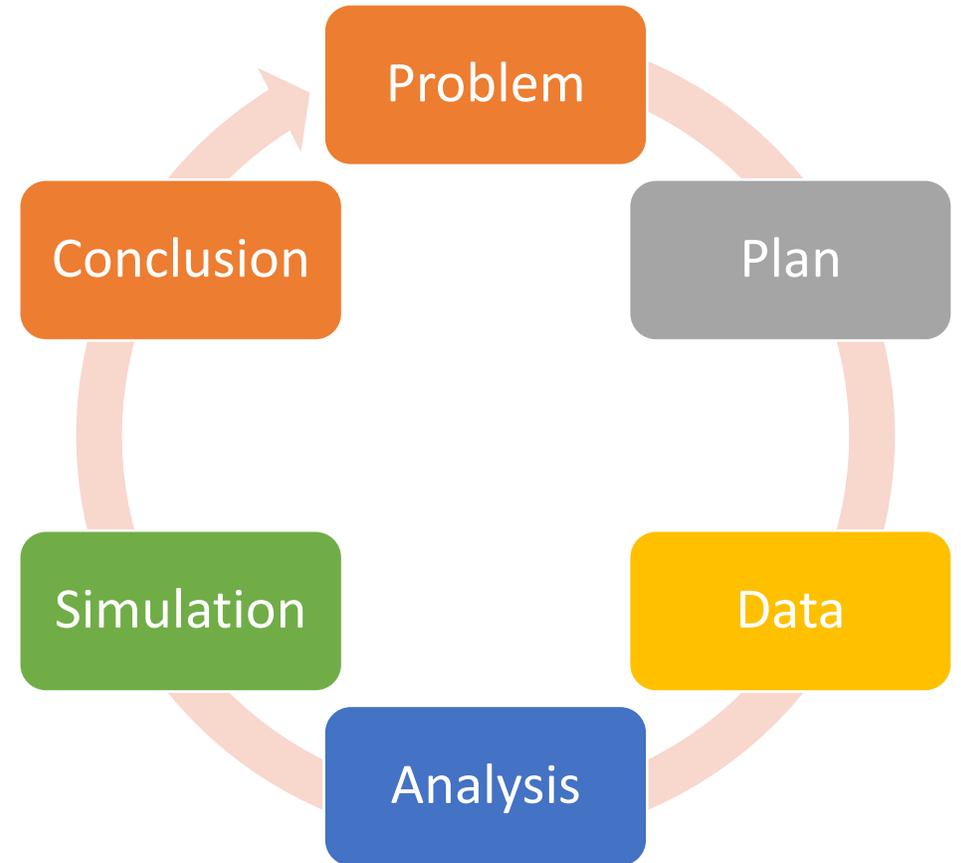
- Data acquisition
- Data collection
- Data analysis
- Interpretation

 PDCA cycle



PPDASC cycle

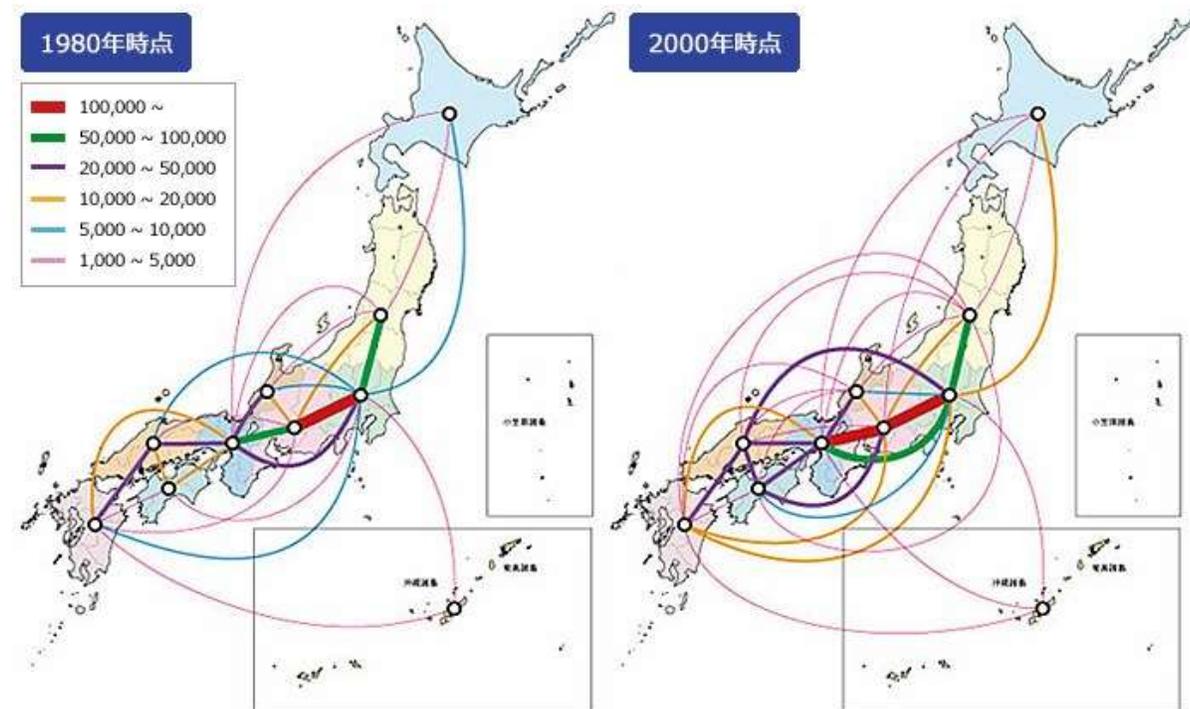
- (1) Start with Small Data acquisition, conduct data analysis and numerical simulation
- (2) Conduct data-driven investigations through the PPDASC cycle
- (3) Scale up the size of data and numerical simulation to approach the final goal of data-driven numerical simulation with Big Data



2地点間の人の移動

- 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な傾向にあるのか？
- 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

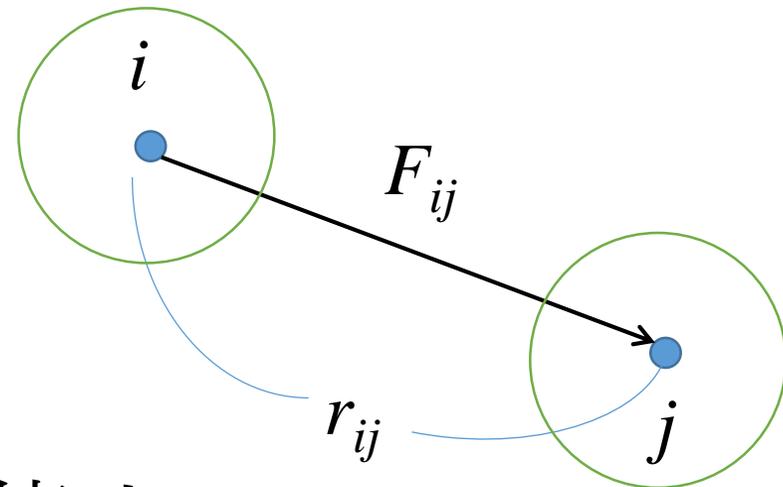
- 2地点間($i \rightarrow j$)の移動量 F_{ij}
- 出発地点*i*の人口 $x_i(d)$
- 到着地点*j*の人口 $x_j(d)$
- *i*と*j*との距離 r_{ij}

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

$x_i(d)$: 場所*i*の*d*[km]近傍人口

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

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



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

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

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

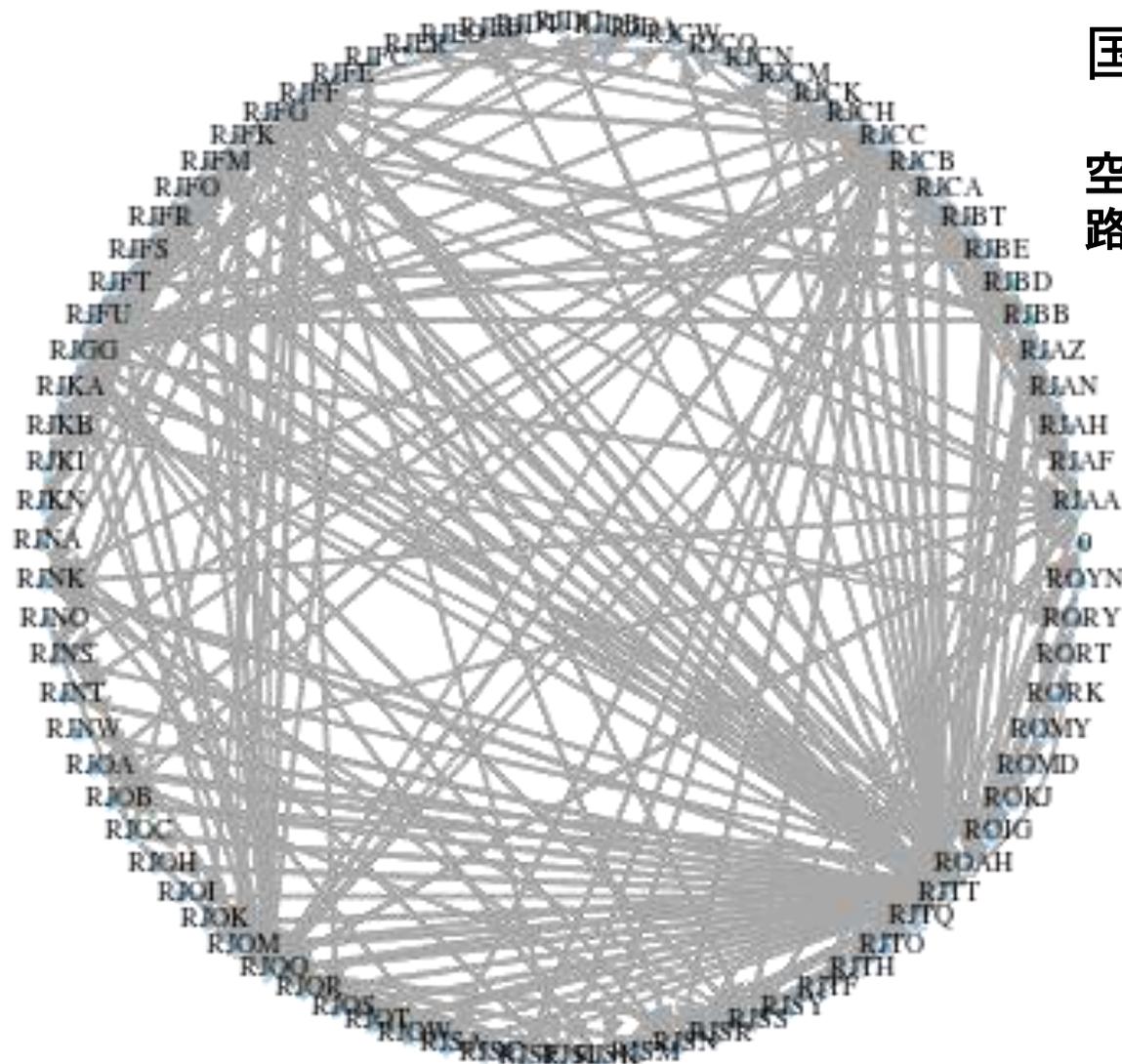
重力モデルの重回帰係数推計

- 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

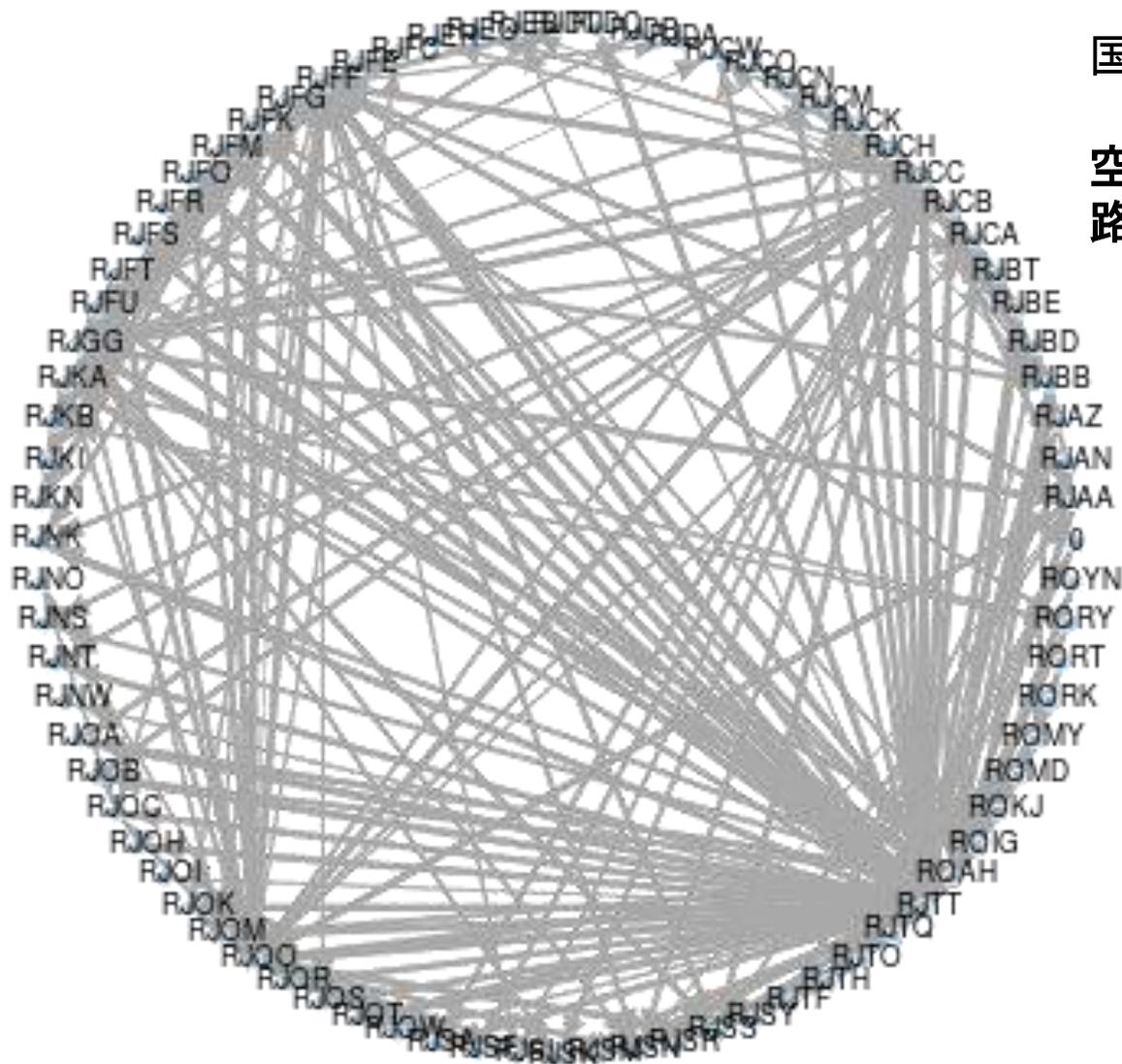


平均出次数 5.372093
平均入次数 5.372093
平均クラスタ係数 0.469923
大域クラスタ係数 0.205842
入次数エントロピ 2.206731
出次数エントロピ 2.206731
次数アソータティビティ -0.436595
平均次数長 2.473598

日本の航空輸送ネットワーク(貨物)

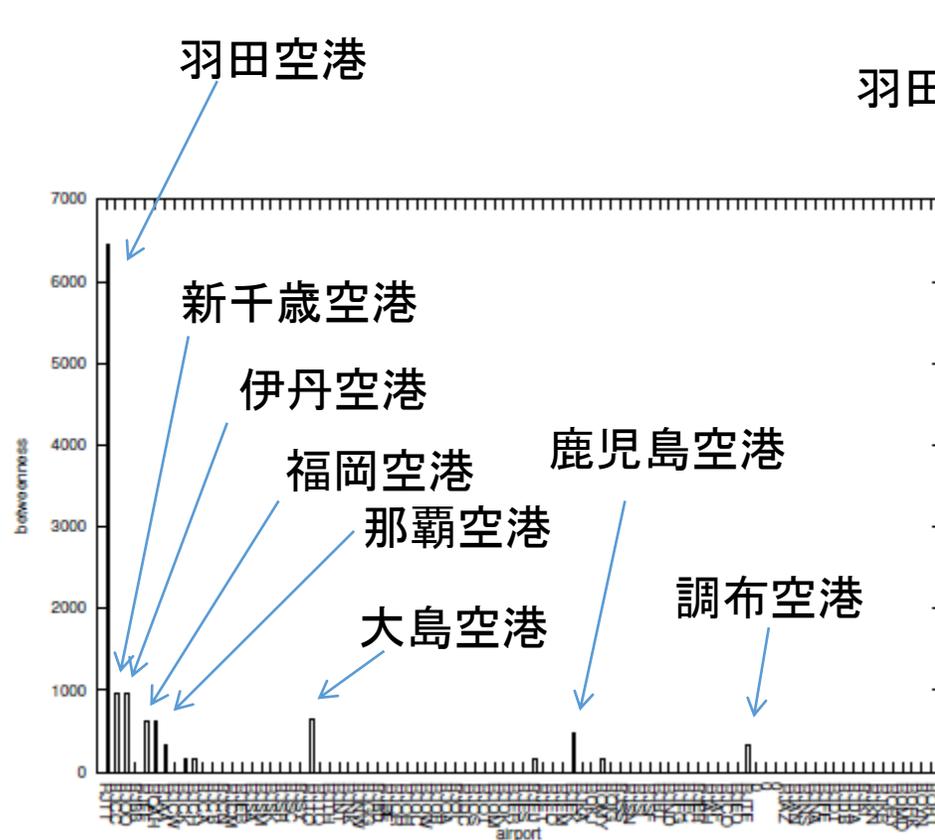
国土交通省2012年航空輸送統計

空港数 86
路線数 476

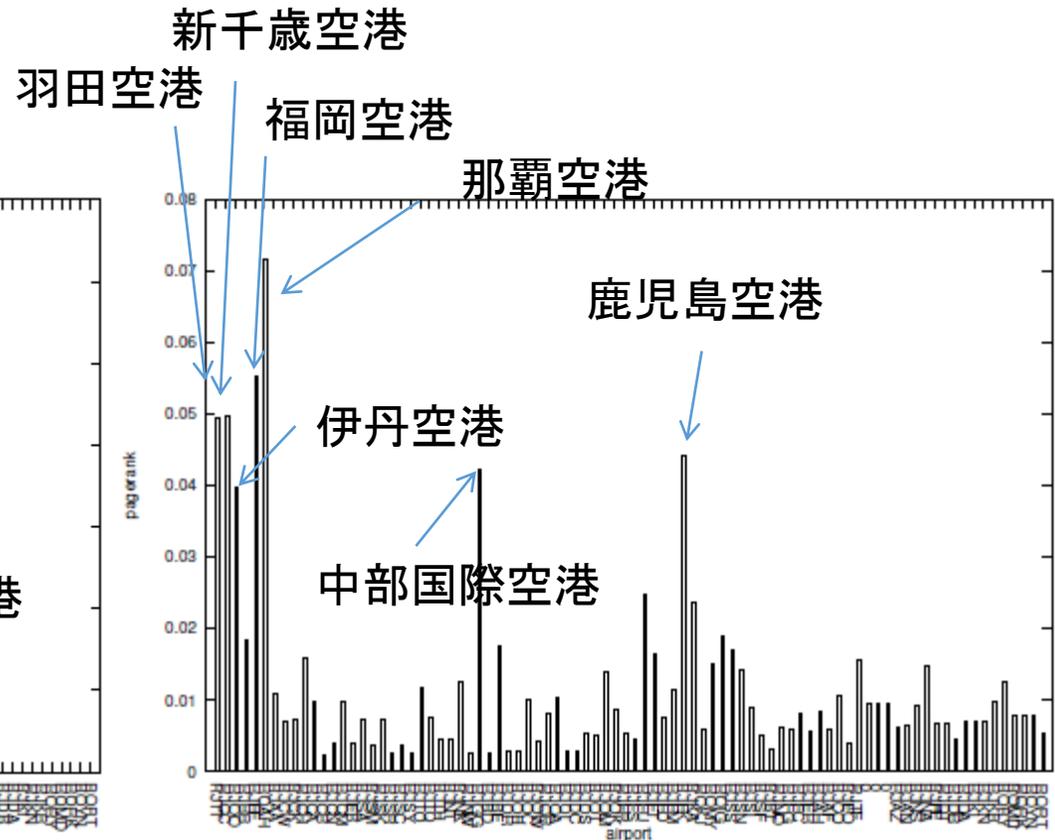


平均出次数 4.804878
平均入次数 4.804878
平均クラスタ係数 0.487389
大域クラスタ係数 0.191510
入次数エントロピ 2.090713
出次数エントロピ 2.090713
次数アソータティビティ -0.468332
平均次数長 2.494731

日本における推計

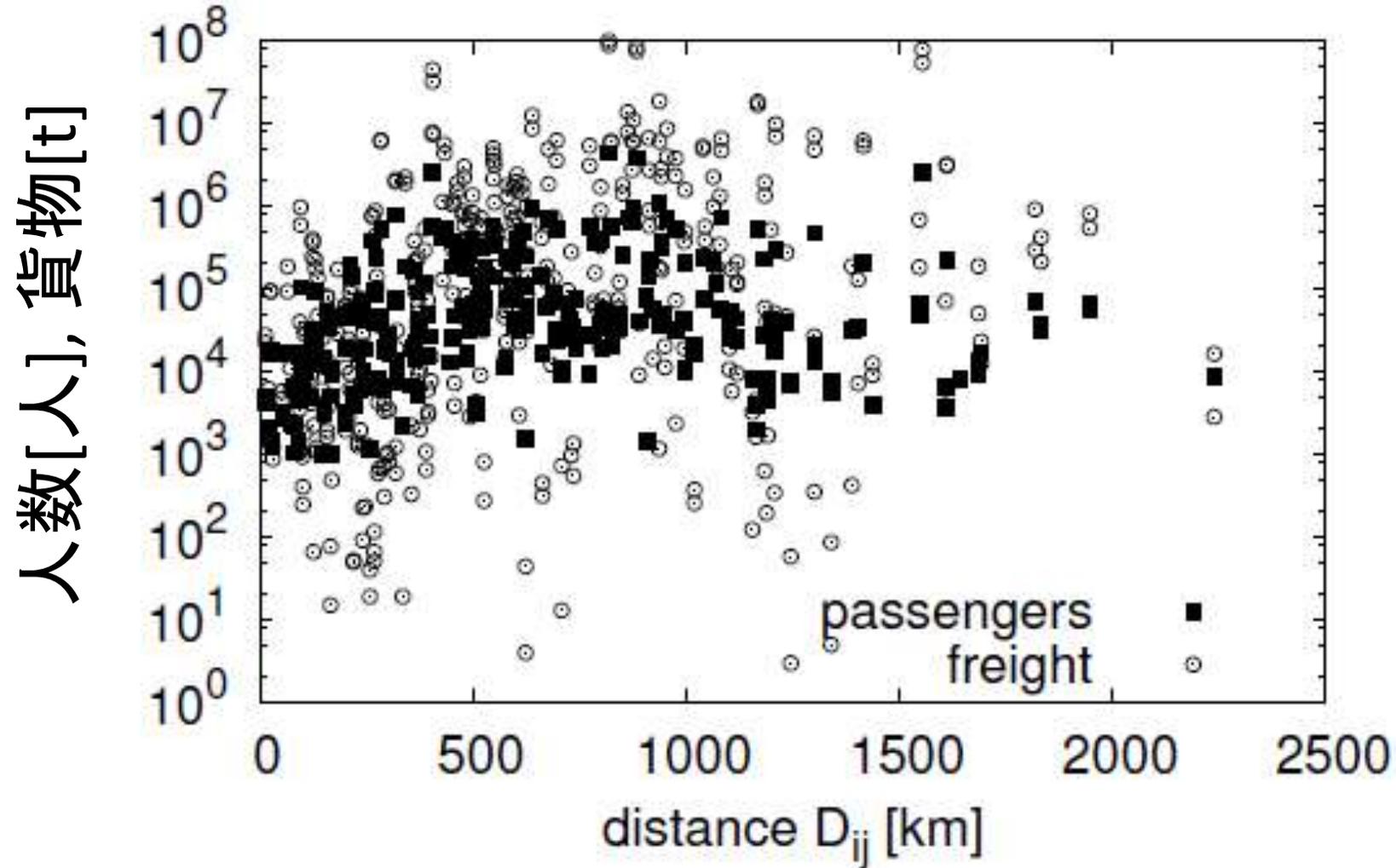


Betweenness centrality



Page Rank

経路距離と移動量・貨物量の関係



総務省統計局 e-Stat

お問い合わせ ヘルプ English 文字拡大・読み上げ

e-Stat
数字で見る日本
e-statは、日本の統計が閲覧できる政府統計ポータルサイトです。
政府統計の総合窓口

統計データを探す 地図や図表で見る 調査項目を調べる 統計サイト検索・リンク集 ログイン

6/9(日)午前1:00~5:00の間、システム作業のためサイトの閲覧ができなくなります。ご迷惑をおかけしますがご了承ください。

アンケート 実施中
ご協力をお願いします

統計について勉強しよう
統計を知る・学ぶ

ランキング

統計キーワード	統計表
利用件数	キーワード
1	220 国勢調査
2	207 人口
3	83 家計調査
4	60 死因
5	59 都道府県
6	53 賃金

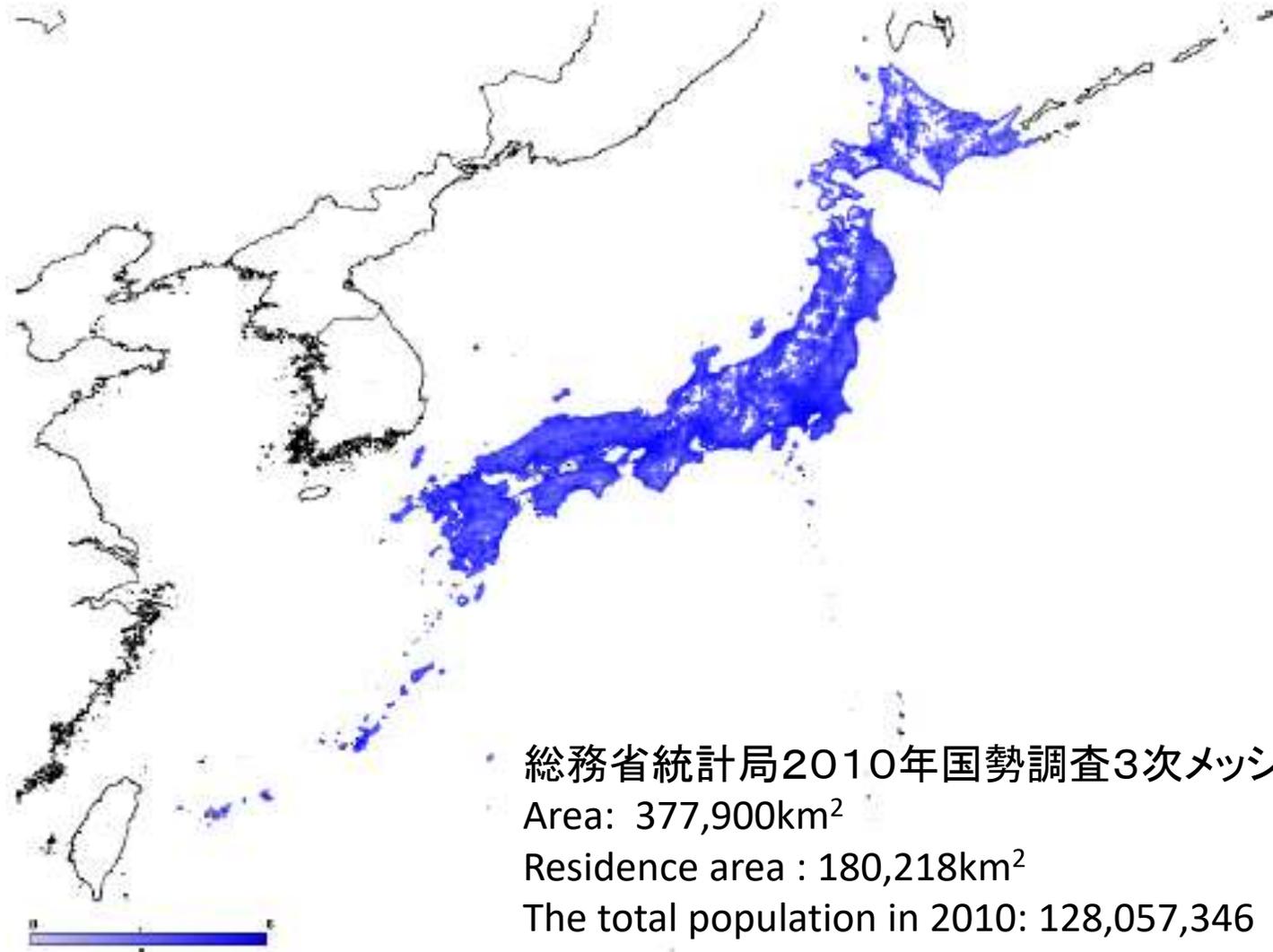
新着情報 公表予定 お知らせ RSSによる配信はこちら

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

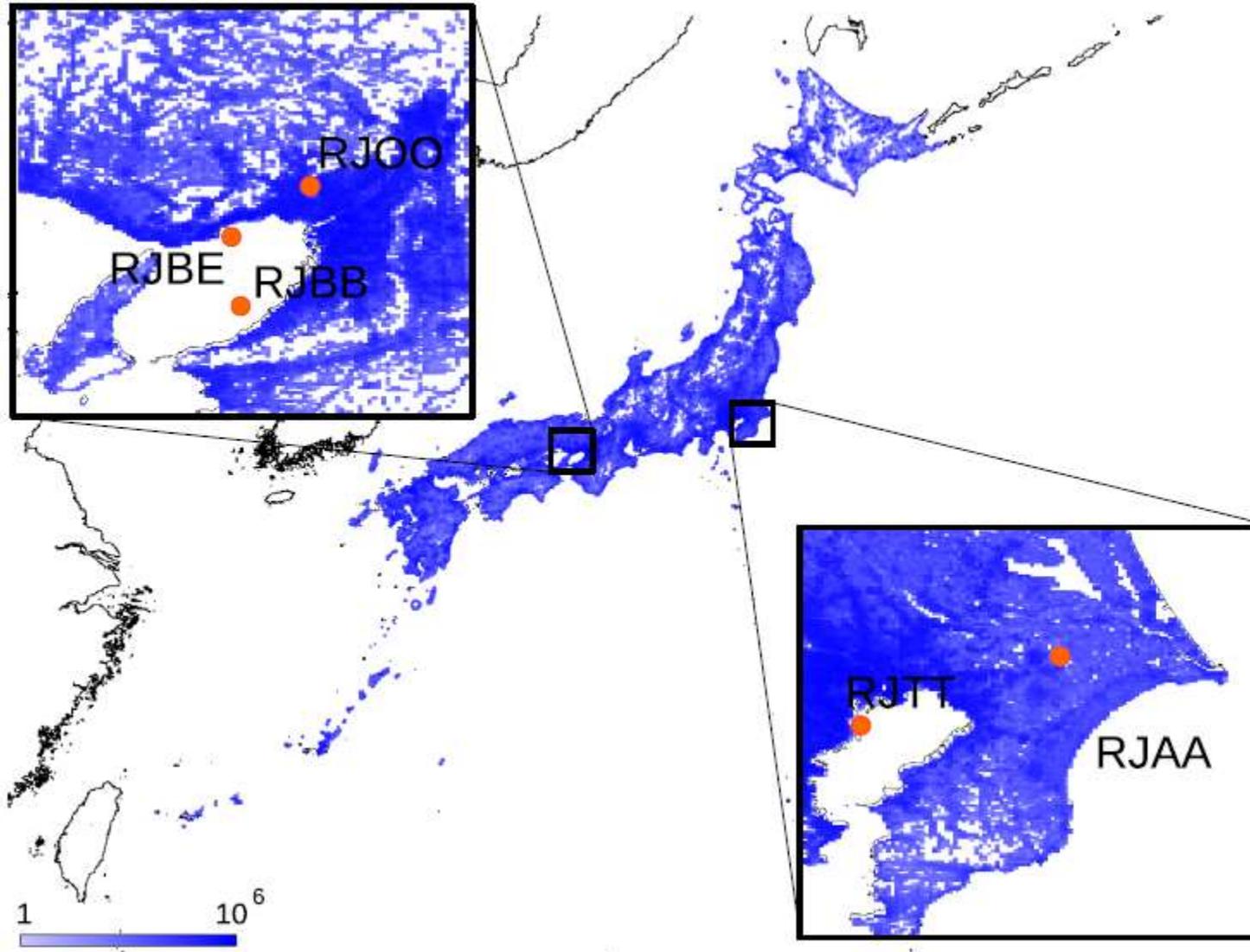
今回利用するメッシュデータ

- 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)

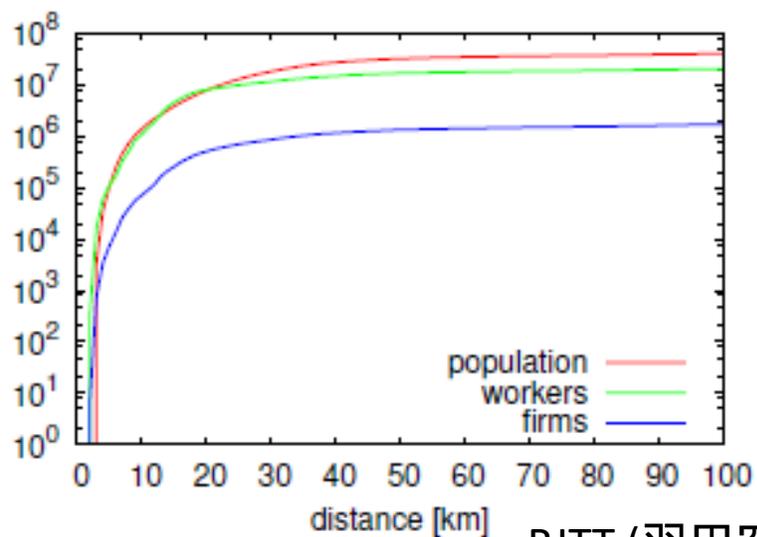
日本の人口



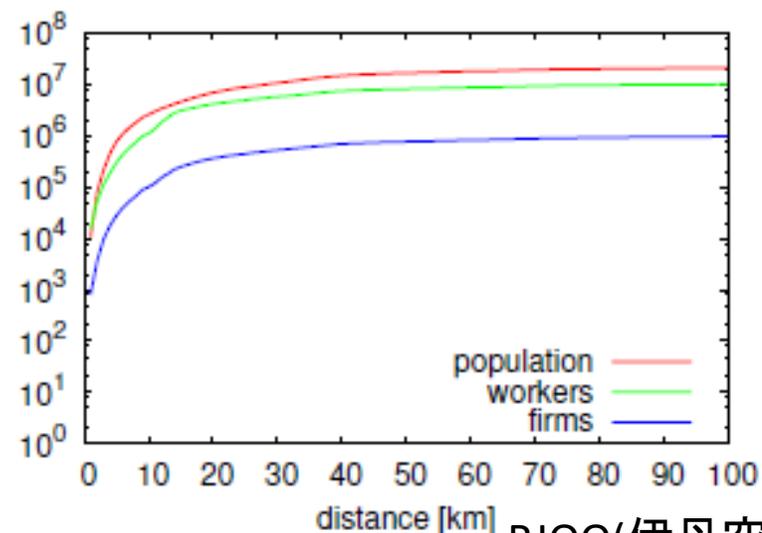
空港周辺の人口



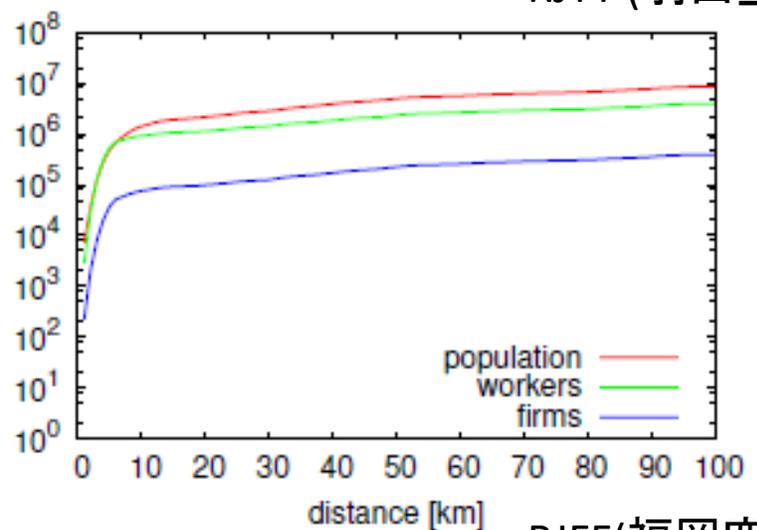
人口、労働者数、会社数



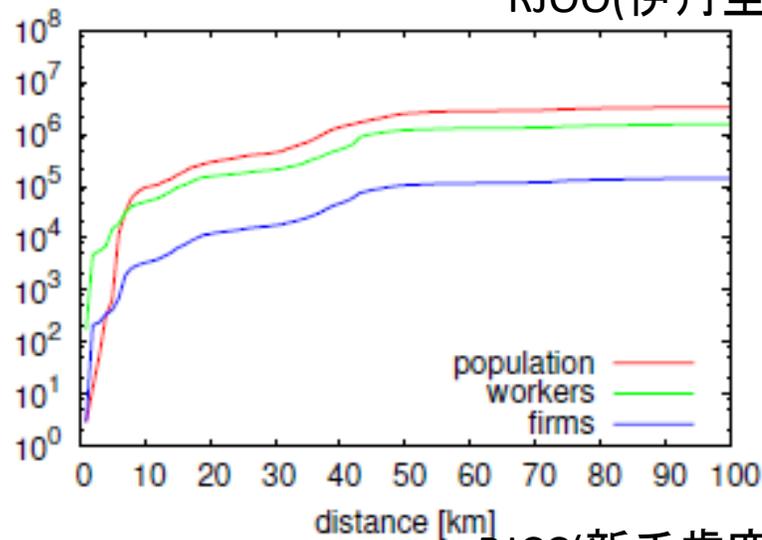
RJTT (羽田空港)



RJOO (伊丹空港)



RJFF (福岡空港)



RJCC (新千歳空港)

正規方程式

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

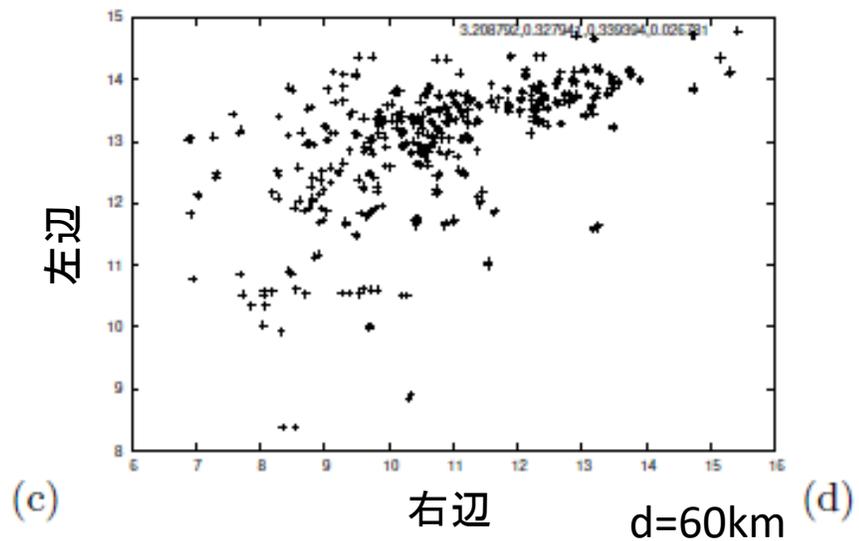
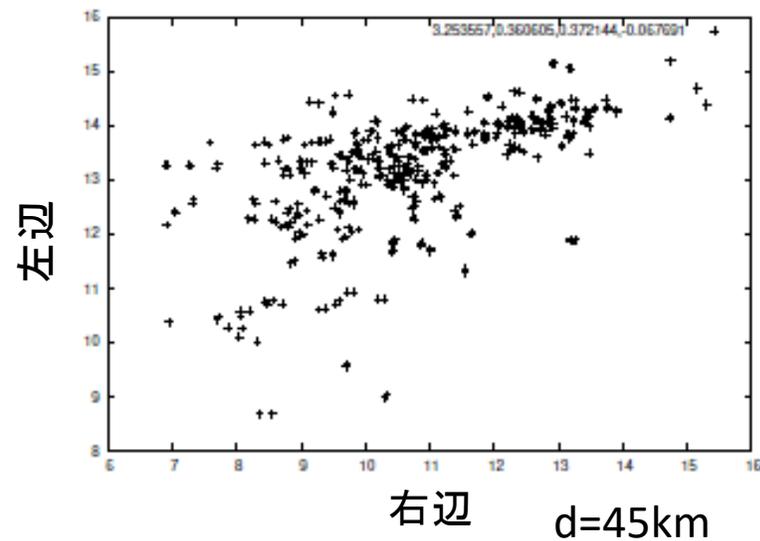
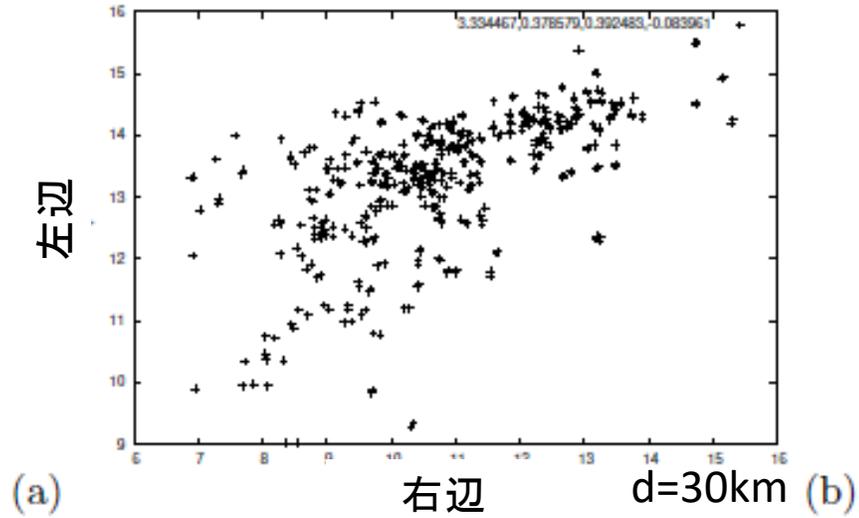
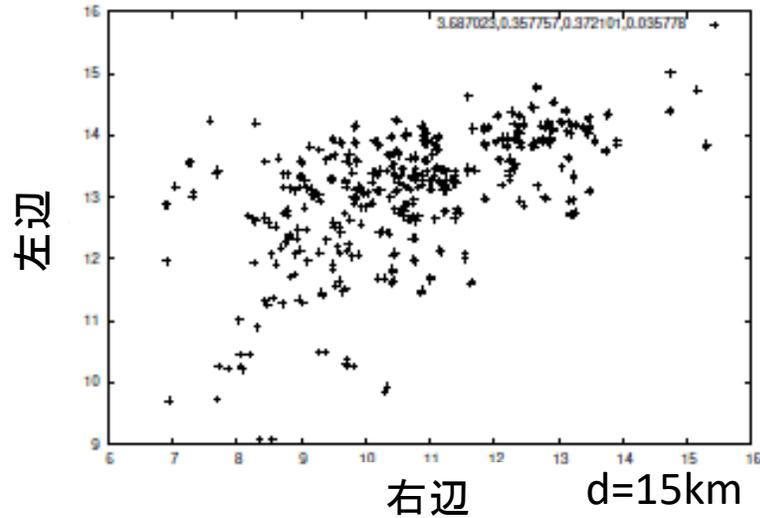
Σ は F_{ij} が0でない全ての空港 i, j についての和を表す
 M は流量データ数

平均二乗誤差

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

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

移動量 F_{ij} と予測値との関係



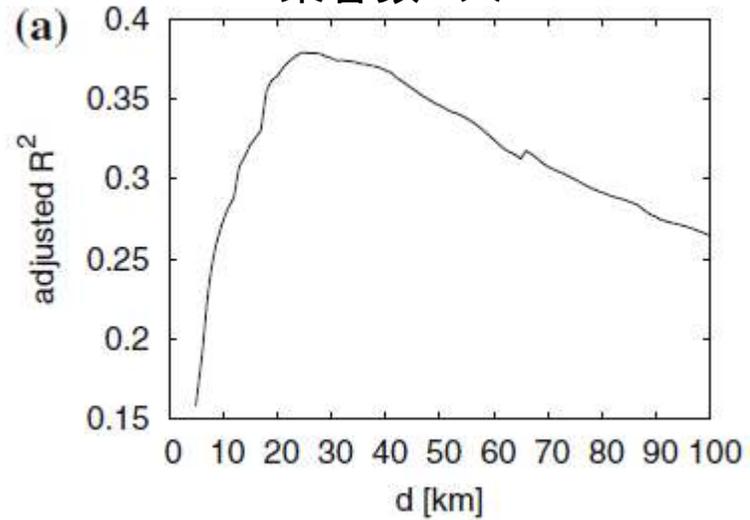
修正済みR²

adjusted $R^2(d)$

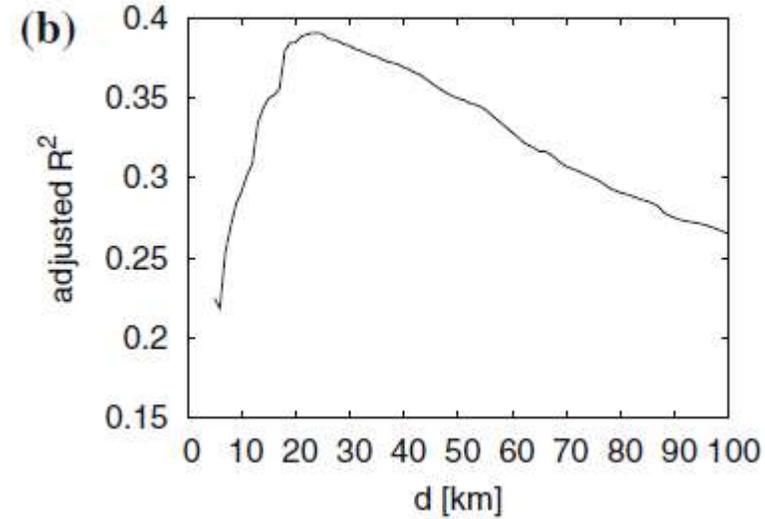
$$\begin{aligned} & \sum_{\substack{i,j \\ F_{ij} \neq 0}} \left(\ln F_{ij} - \beta_0(d) - \beta_1(d) \ln x_i(d) - \beta_2 \ln x_j(d) - \beta(d) \ln r_{ij} \right)^2 / (M - 5) \\ = & 1 - \frac{\sum_{\substack{i,j \\ F_{ij} \neq 0}} \left(\ln F_{ij} - \frac{1}{M} \sum_{\substack{i,j \\ F_{ij} \neq 0}}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

修正済みR²

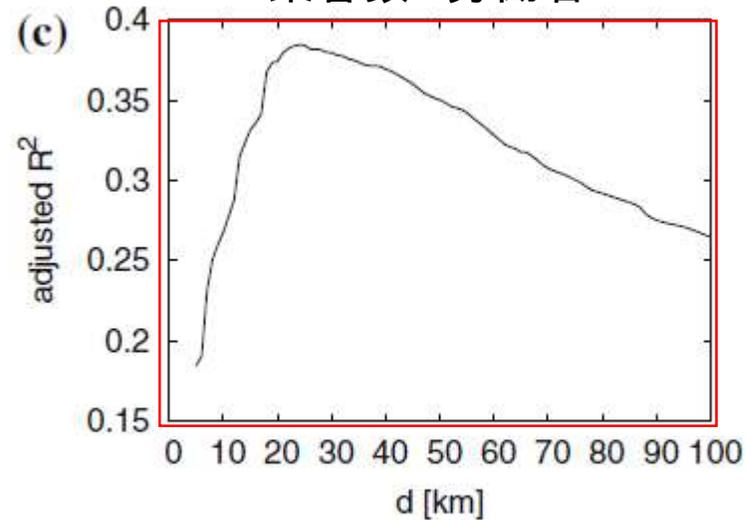
乗客数~人口



乗客数~事業所



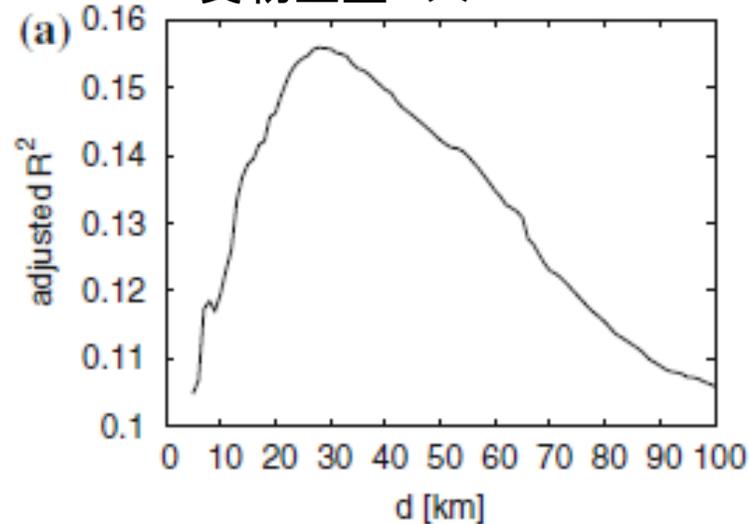
乗客数~労働者



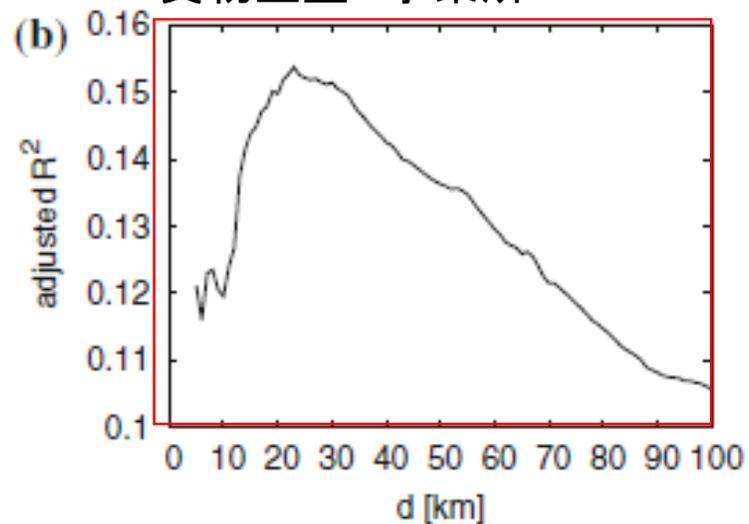
MSE	$\beta_0(d)$	$\beta_1(d)$	$\beta_2(d)$	$\beta_3(d)$
1.314442 (0.389791)	3.001870	0.311012	0.316914	-0.019794
(std.)	0.493901	0.028937	0.028937	0.086524
(t-val)	6.077873	10.74788	10.95185	-0.228768
(Pr[> t])	0.000000	0.000000	0.000000	0.819152

修正済みR²

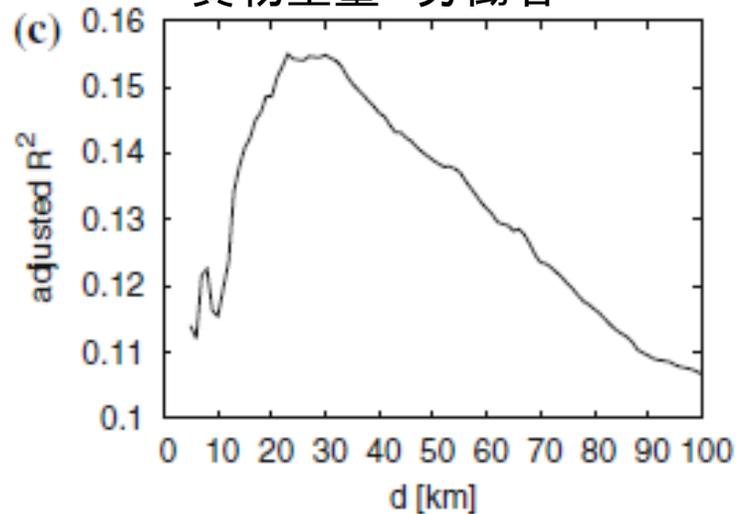
貨物重量 ~ 人口



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

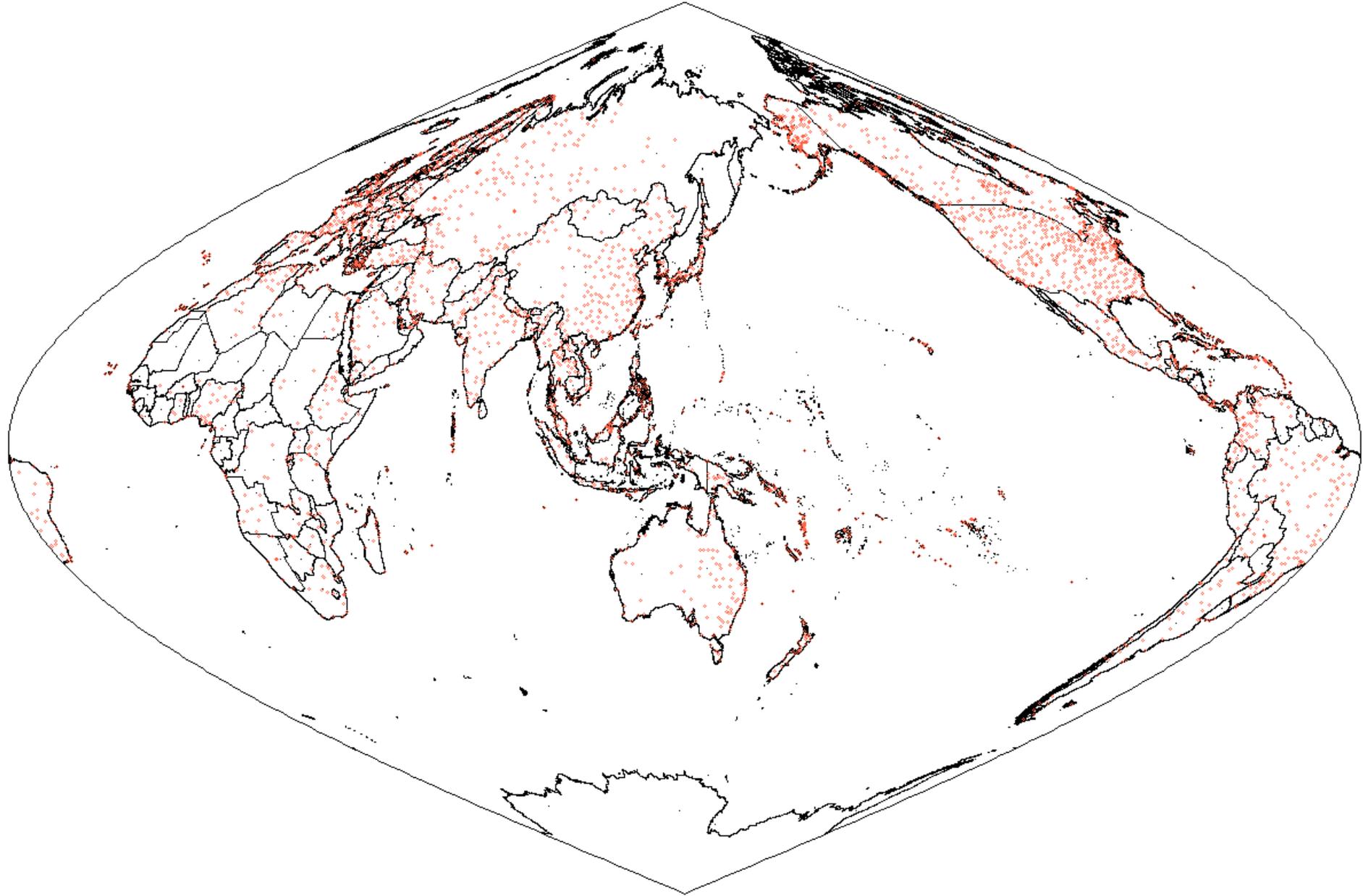
例えば

空港	ICAO	緯度	経度	労働者数 (25km)	事業所数 (23km)
長崎	RJFU	32.916906348	129.914001465	340016	32441
熊本	RJFT	32.8372993469	130.8549957275	500020	44420
新千歳	RJCC	42.7751998901	141.6920013428	176685	13651
青森	RJSA	40.7346992493	140.6909942627	255034	22317



Departure	Arrival	Geodesic distance (km)	Passengers (persons)	Freight (Mg)
RJSA	RJFU	1292.53	9468.27	42870.21
RJSA	RJFT	1238.26	10461.30	47901.91
RJCC	RJFU	1503.50	8487.00	38814.69
RJCC	RJFT	1455.15	9375.22	43552.31

主要な民間空港



2014年

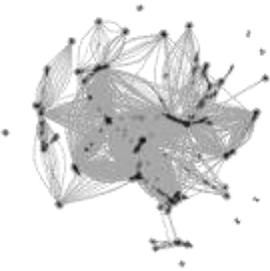
Aki-Hiro Sato, Hidefumi Sawai , “Risk Assessment for a Global Air Transport System Using Socioeconomic-Technological-Environmental Databases,” 2016 IEEE 40th Annual Computer Software and Applications Conference (2016) pp. 572-581

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

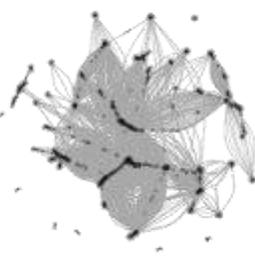
mean in-degree	13.80722
mean out-degree	13.80722
average path length	4.031476
assortativity	0.003953
Average betweenness weighted by seat	36044.5 (seats)

世界航空輸送ボリュームデータ

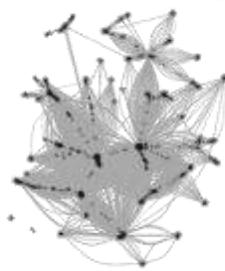
Jan. 2014



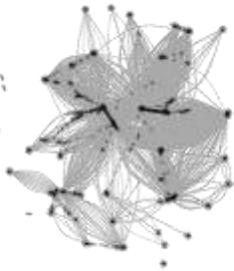
Feb. 2014



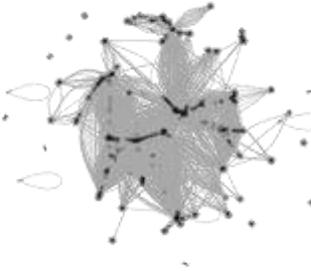
Mar. 2014



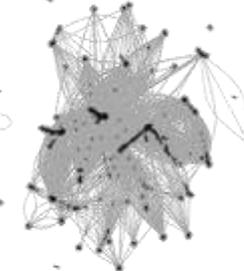
Apr. 2014



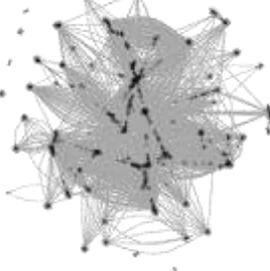
May. 2014



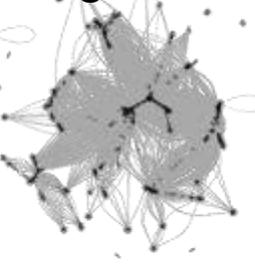
June 2014



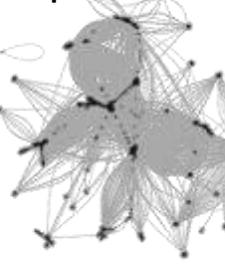
Jul. 2014



Aug. 2014



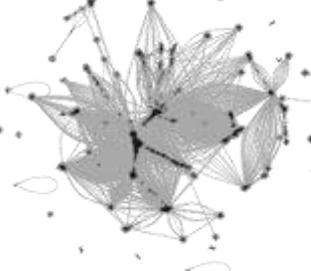
Sep. 2014



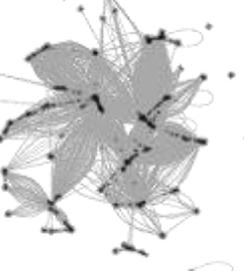
Oct. 2014



Nov. 2014



Dec. 2014



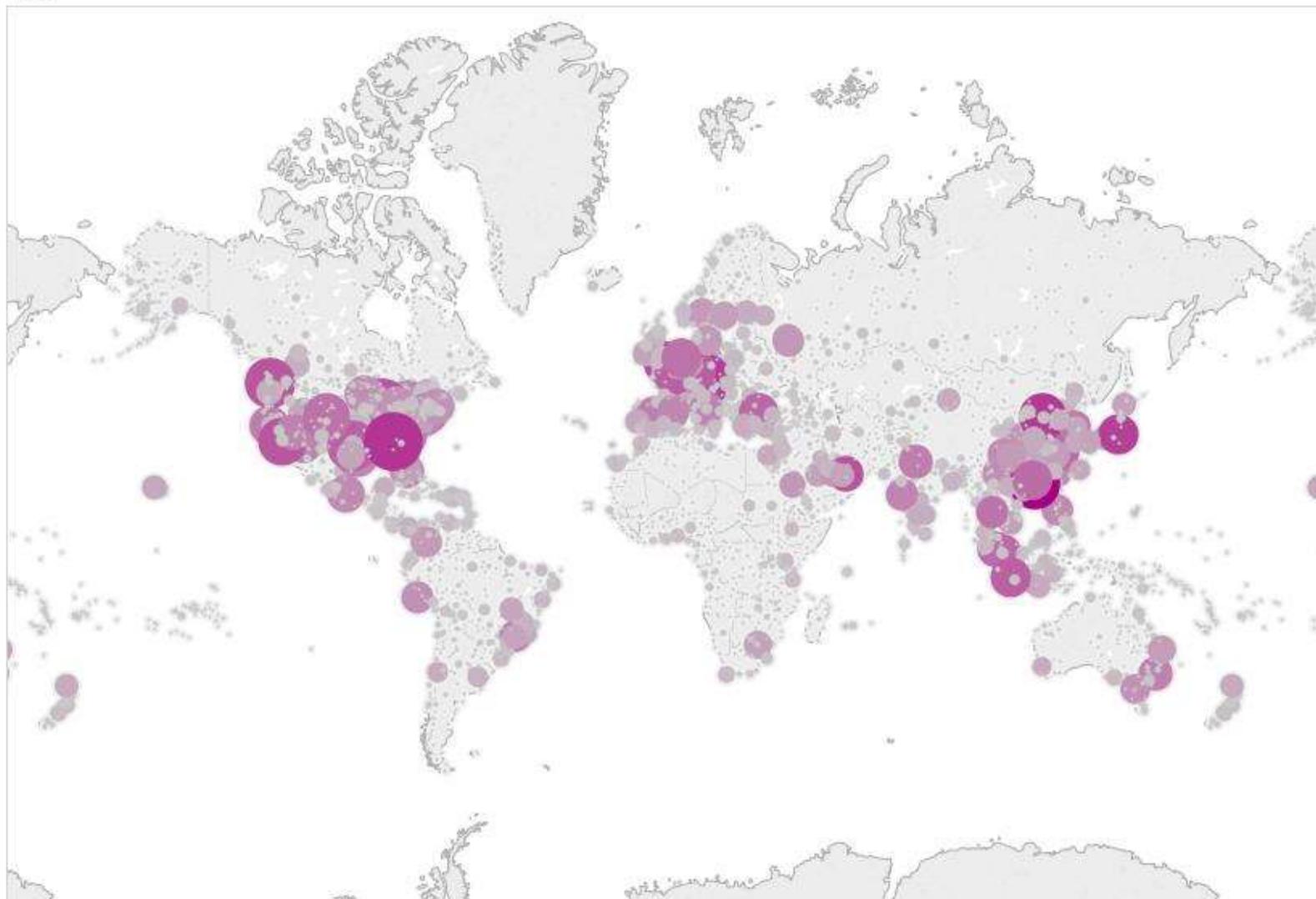
世界航空輸送ボリュームデータ

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
○の大きさ: フライト数

freq



集計(SUM([depseats])+[.])
9 160,745,626

発着数と座席数(上位10位)

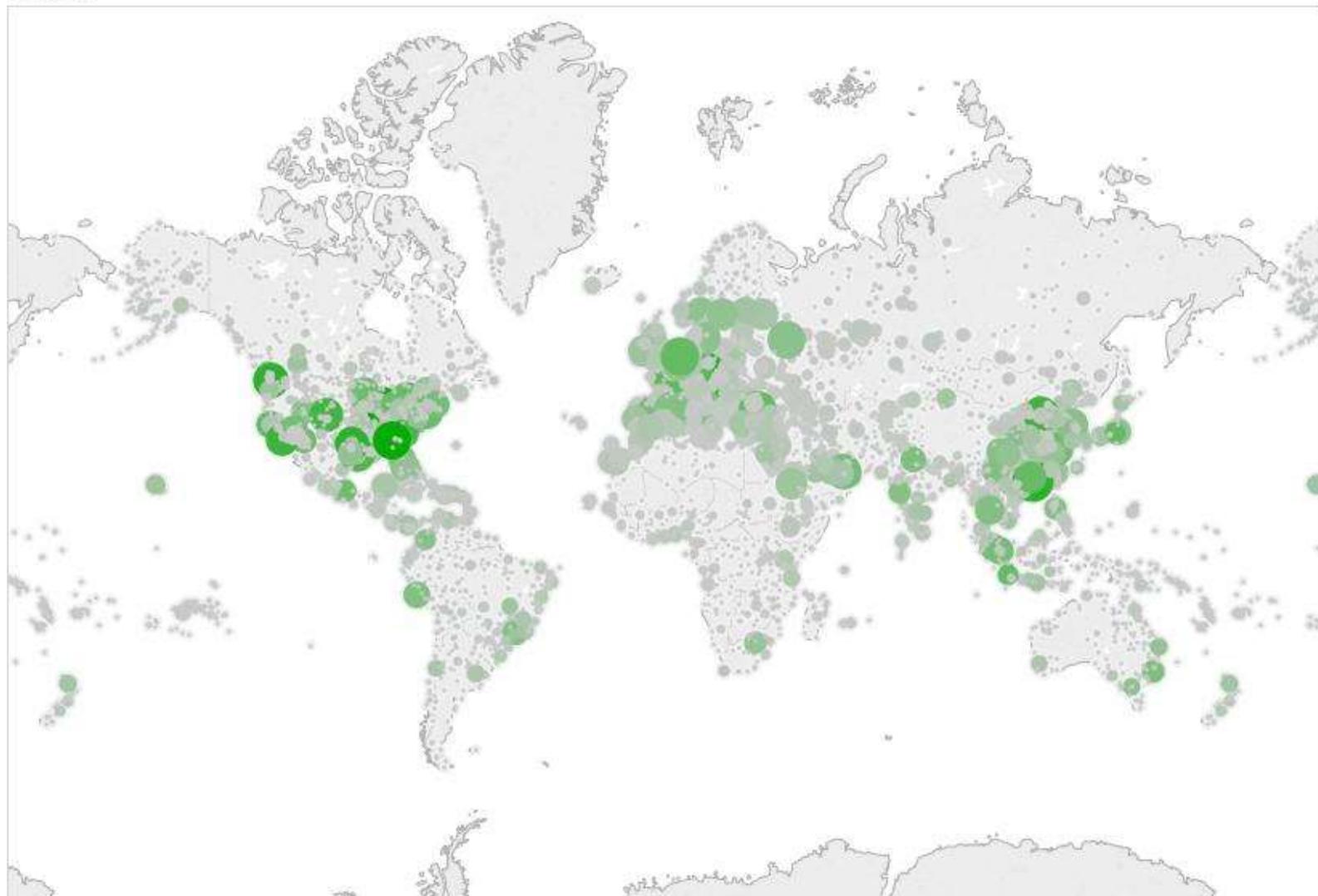
IATA	発着数
ORD	879,754
ATL	850,532
MUC	707,362
DFW	661,066
HKG	656,488
SEA	647,878
LAX	585,608
PEK	568,564
DEN	551,672
CLT	513,991

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

次数中心性

色の濃さ:フライト数
○の大きさ:次数中心性

degree



SUM([depfreq]+[arrfreq])



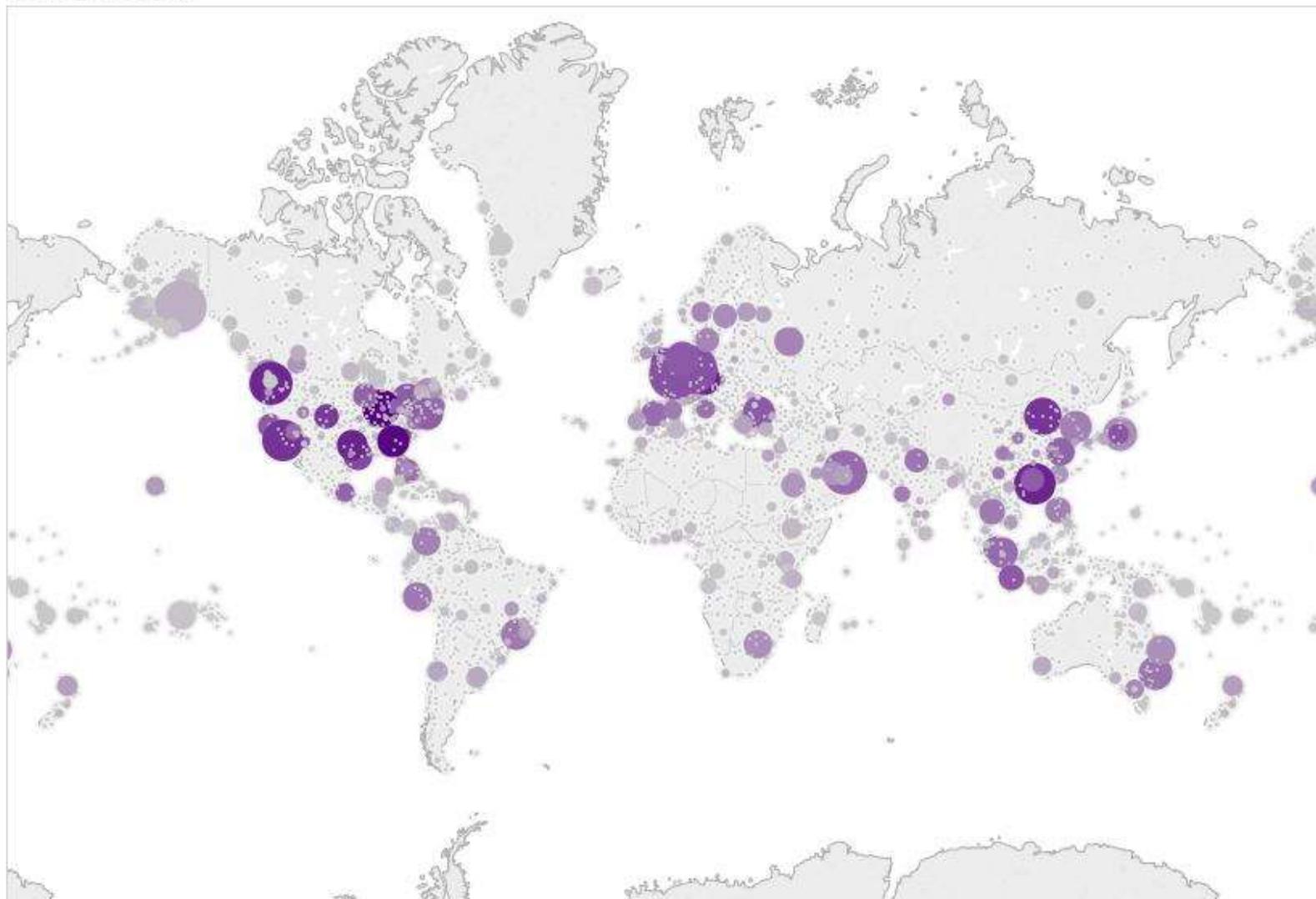
1

879,754

媒介中心性

色の濃さ:フライト数
○の大きさ:次数中心性

betweenness_bi

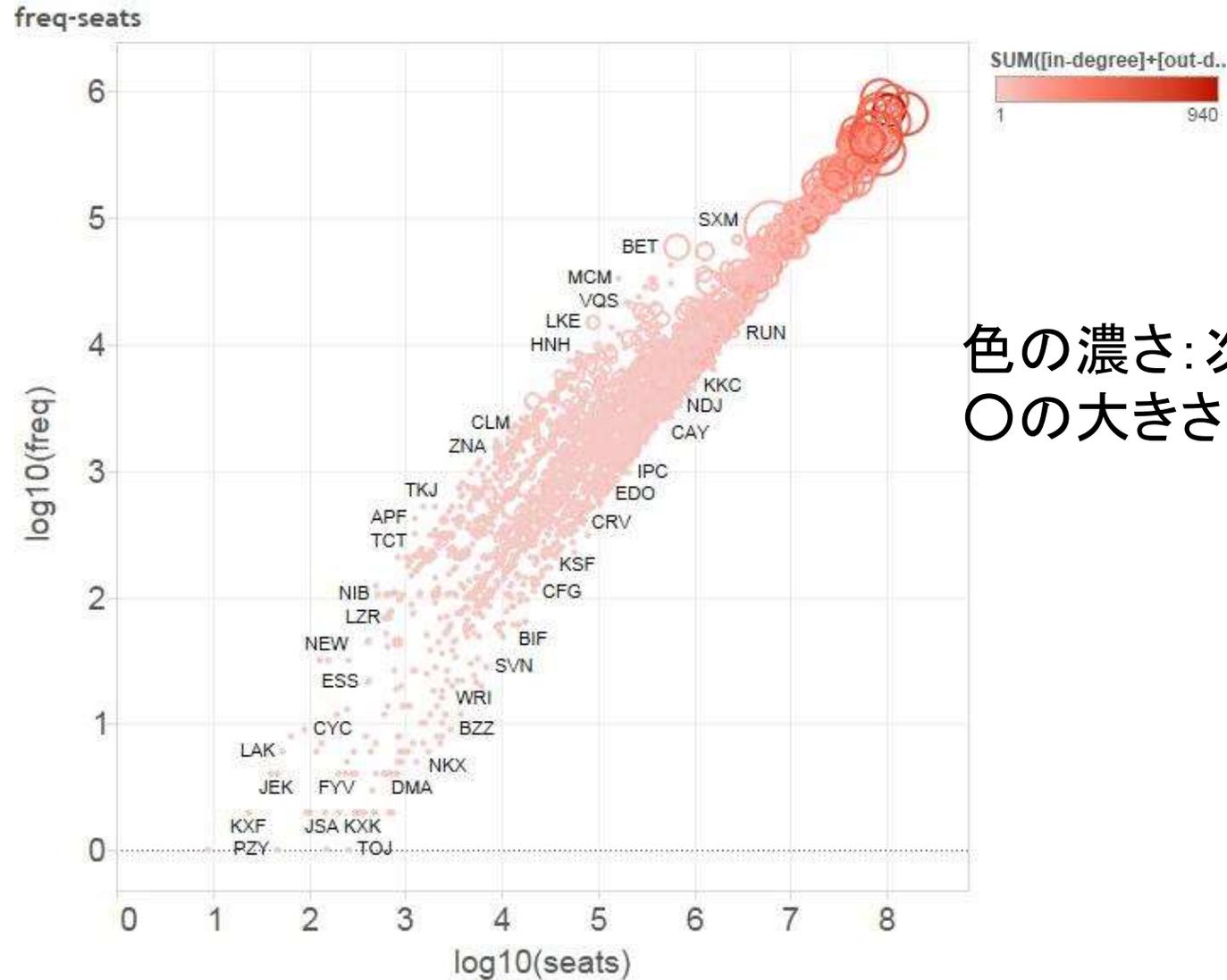


次数中心性と媒介中心性 (上位10位)

IATA	次数
MUC	940
HKG	606
FRA	594
CDG	590
AMS	528
ORD	514
ATL	510
IST	499
DME	487
PEK	464

IATA	媒介中心性
ANC	1,178,448.136
CDG	1,102,087.635
FRA	1,006,040.152
DXB	857,797.233
SEA	810,619.172
LAX	754,122.032
HKG	748,079.229
PEK	611,498.478
ORD	608,469.023
AMS	518,815.127

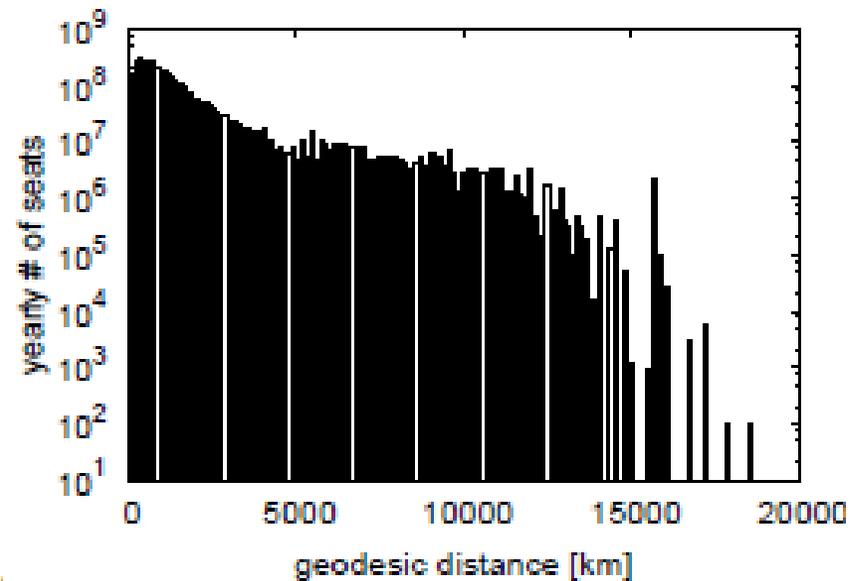
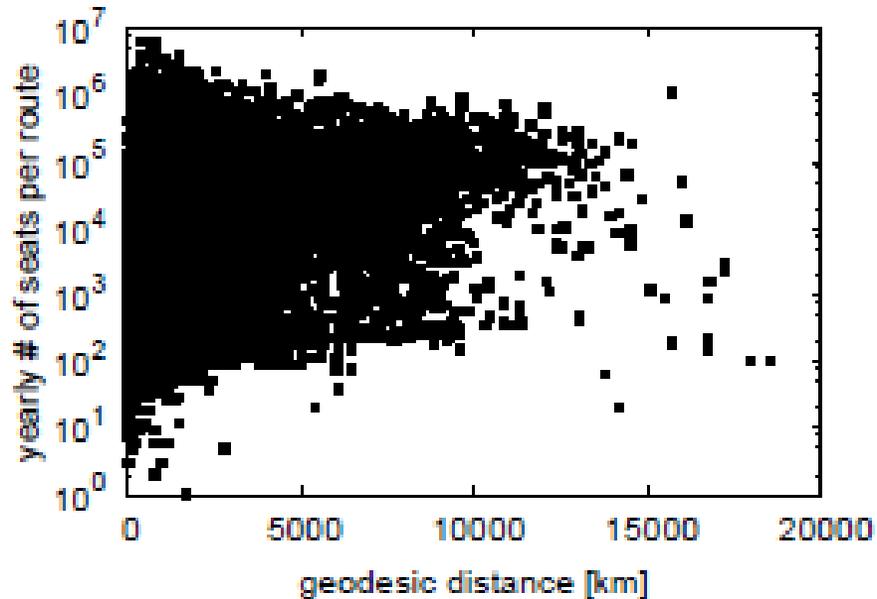
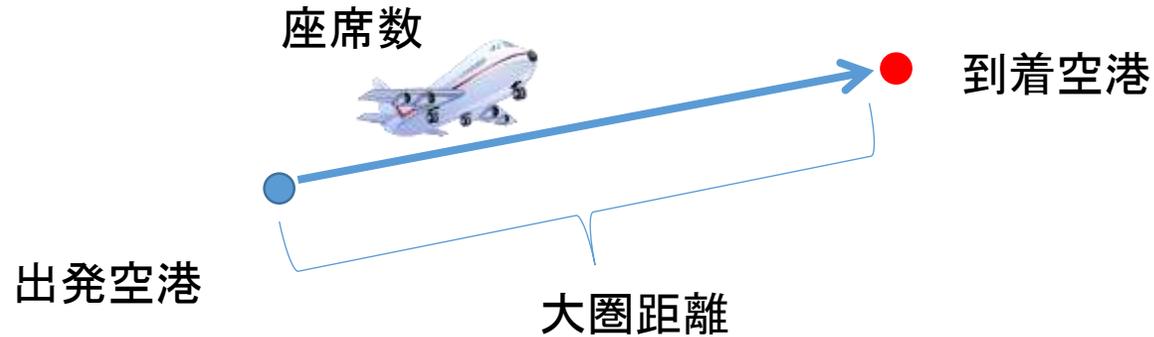
座席数とフライト数との関係



色の濃さ: 次数

○の大きさ: 媒介中心性

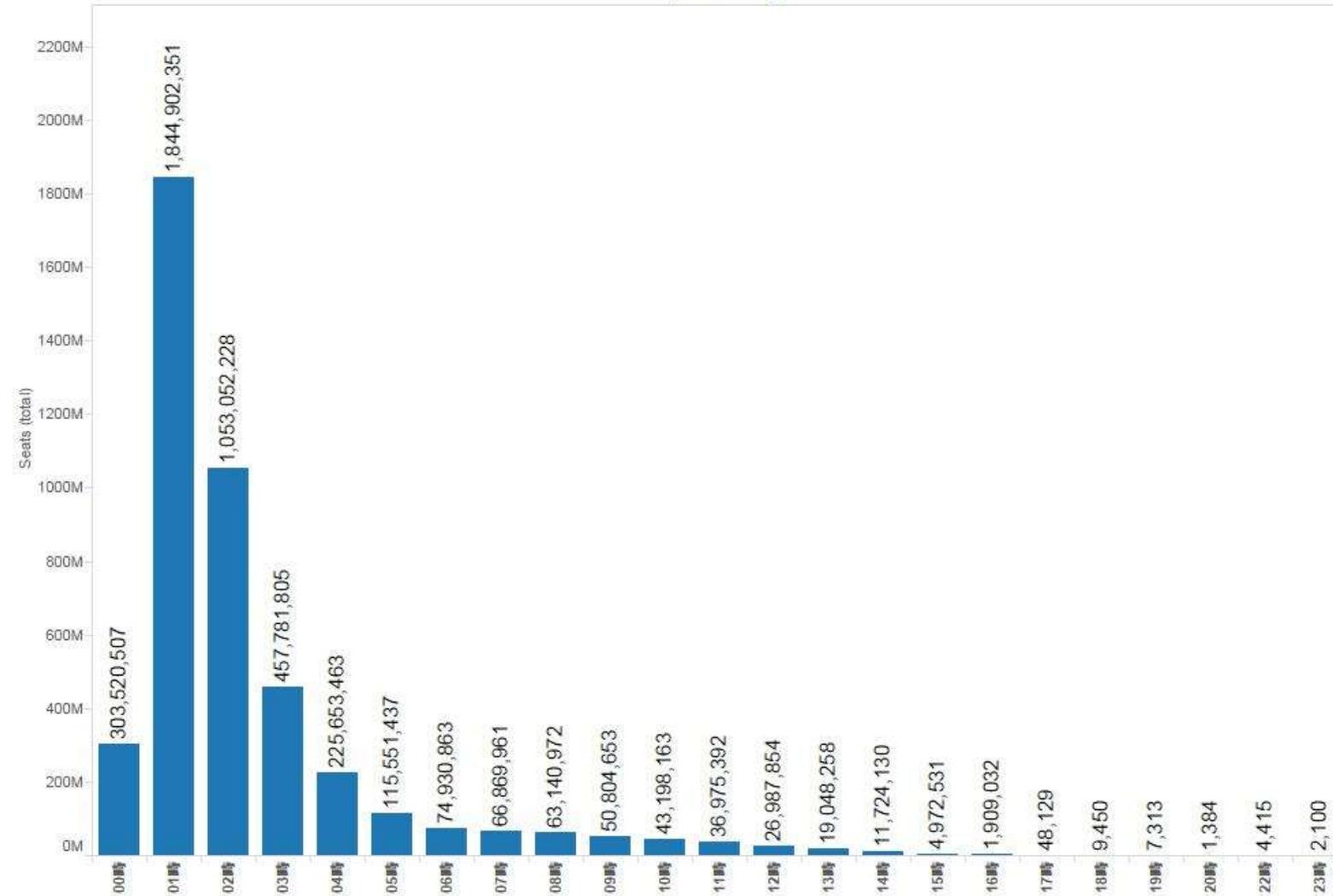
空港間距離と座席数との関係



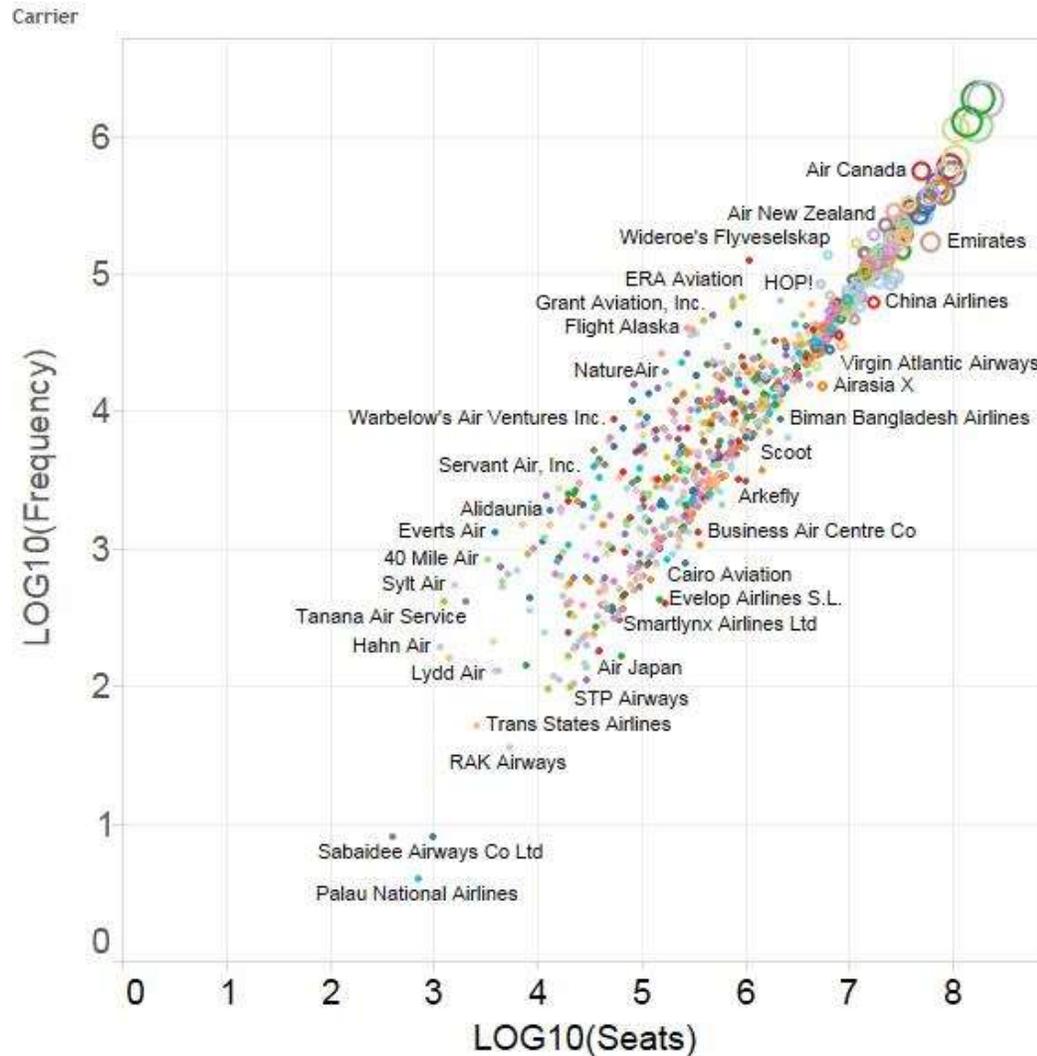
飛行時間と座席数

ElapsedTime

ElapsedTime の時間



航空会社ごとの座席数とフライト数



色: 航空会社

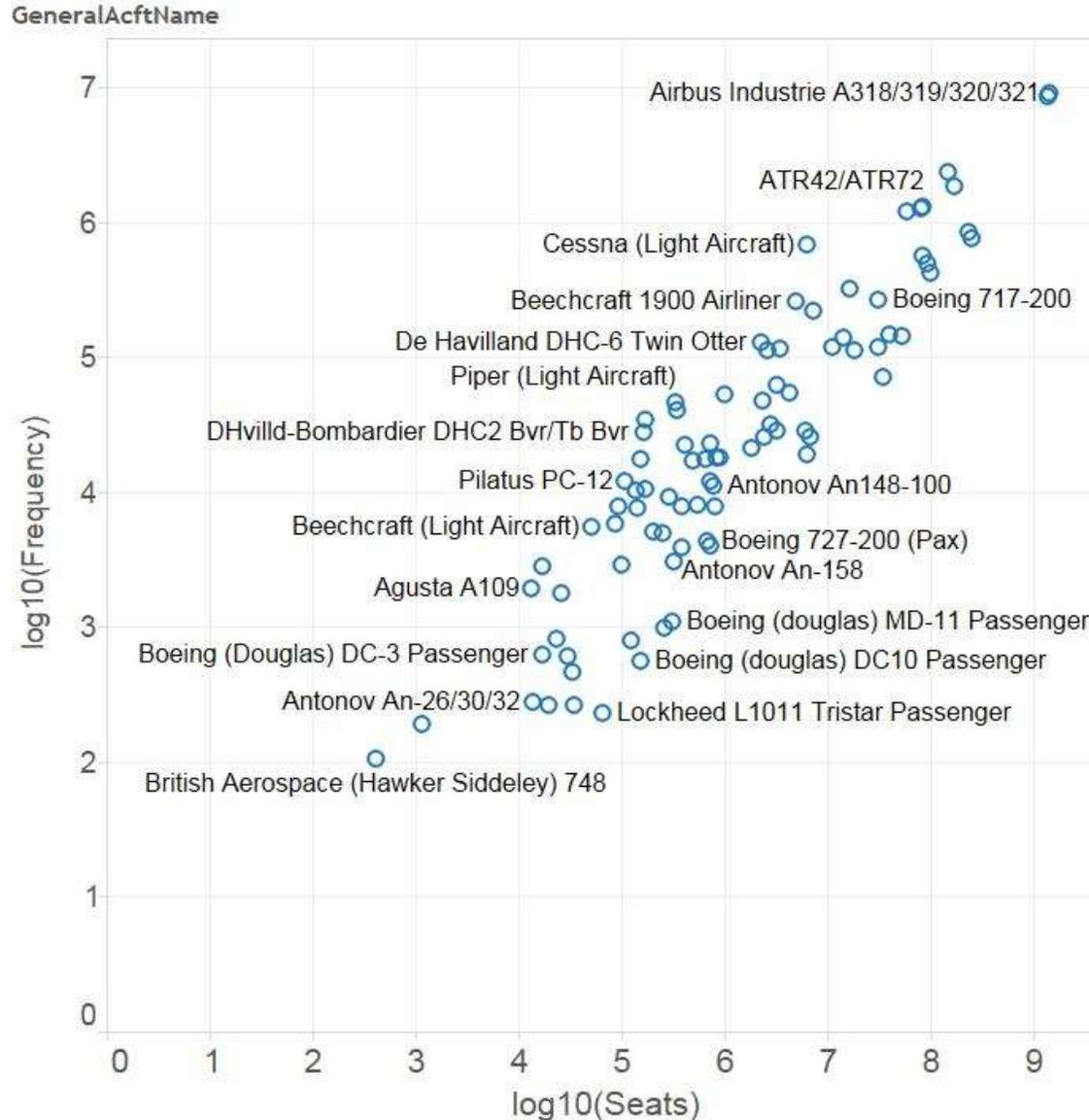
○の大きさ: フライト数

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

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