地方創生データウェアハウス *JapanReview.Com* の構築とその活用可能性

JapanReview.Com The development of a data warehouse for local-level analysis and its potential for revitalizing local economy

岡本 悦司　　神谷 達夫

要旨

市町村単位の詳細なデータが e-STAT 等で公表されるようになり，地域の特性や実態を把握することによって地方創生に活用できると期待される。しかしながら，データが膨大であることから，その活用には技術的困難が伴う。膨大な統計表データをウェブ上で Excel のピボットテーブルのように自在に活用できるデータウェアハウスを構築したので，その概要を紹介するとともに，活用例を示す。

キーワード: データウェアハウス，e-STAT，オープンデータ，キューブ化，市町村データ

Keywords: data warehouse, e-STAT, open data, cube, municipal-level data

１．データウェアハウスとは

データは通常クロス表[行と列に次元，真中にデータがある]の形式で提供される。しかしそのままでは自由な処理はできない。ひとつひとつのデータに行と列の次元をつけて縦長にした形式にすれば，Excel 上ならピボットテーブルという機能を用いて自由に加工できる。それはあたかも，ルービック・キューブのように次元を自由に動かせることから「キューブ」形式と呼ばれる(図 1)。

Excel ではキューブ形式のデータからピボットテーブルを作成できる機能に加えて，逆にクロス表をピボットテーブルに変換し，さらにはキューブ形式に変換する「逆ピボットテーブル」という機能も備わっており，加工には専らこの技術が用いられる。

表記の「揺れ」統一も DWH 作成上重要な作業である。総務省は自治体ごとに５ケタコードを振っているが，コードだけでは自治体名がわからないので DWH では「コード＋都道府県＋市町村」に統

一した(町村については郡は省略)。

たとえば茨城県の龍ケ崎市は旧字体の「龍」大文字の「ケ」が正式だが，実際には「龍ヶ崎」「竜ケ崎」「竜ヶ崎」と様々な表記があり，統一されないと異なる統計のデータを市町村単位で結合できない（「ケ」にはさらにケ,ヶ,ヶと 3 種の字体がある）。DWH 作成にあたっては異なる統計調査で表記が異なっても「08208 茨城県龍ケ崎市」に統一した(「ケ」は大文字の「ケ」に統一)。市町村コードを振ったのは，町村では同一都道府県内に複数あることがあり[たとえば群馬県にはかつて東村が５つも存在した]，さらに DWH を表示させたら時に必ず決まった順番に並ばせることでみやすくするためである。同様の表記の「揺れ」は塩竈市と塩竈市，平塚市と平塚市，聖籠町と聖篭町，諫早市と諫早市，砺波市と礪波市，南砺市と南礪市，鯵ケ沢町と鯵ケ沢町等でもみられる。甚だしい場合は，南大隅町を南大隈町とする誤記も公的統計においてさえみられた。

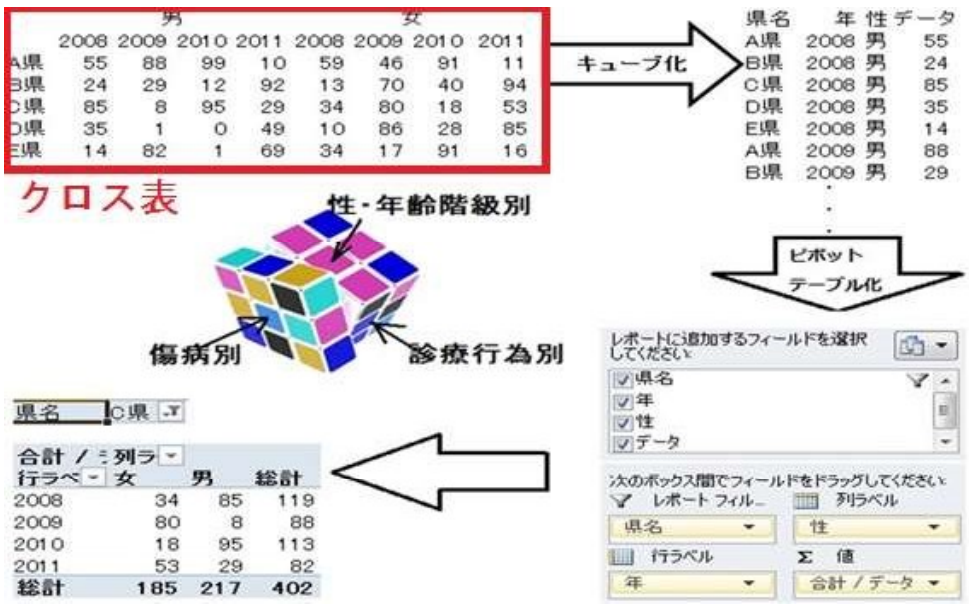

図 1　クロス表のキューブ化とピボットテーブル処理との関連

2. ウェブ上での公開

ピボットテーブル機能は，Excel のあるパソコンでなければ使用できないが，理想的にはウェブ上で自在に活用できれば便利である。オープンソースの Javascript を用いてウェブ上でさながら Excel ピボットテーブルのように操作できるようにして公開した。URL は以下の通りである。

http://www.japanreview.com

収録済データは本稿執筆時点では以下の通りであり，順次拡大してゆく。

総合統計

●市区町村の指標[115MB]

●住民基本台帳人口[100MB]

●将来推計人口[56MB]

産業統計

●農作物統計[市町村単位 17.4MB]

●集落営農統計[市町村単位 5MB]【2015 年度のみ】

●工業統計[市町村単位 56MB]

●建築着工統計[市区町村単位 70MB]

税務統計

__国税[税務署単位]

●源泉所得税[1.9MB]

●申告所得税[51MB]

●酒税[1.8MB]

__住民税[市町村単位]

●市町村民税課税状況【主たる所得別】[51MB]

●住民税所得割課税状況【所得種類別】[18MB]

●軽自動車税[1.8MB]

3. DWH の使用法

DWH の使用法を「福知山市の 5 年間の農作物別の収穫量(t)を表示させる」を例として説明する。

3.1 データのダウンロード

左ウィンドウより表示させたい DWH をクリックする。するとデータのダウンロードが開始され，画面に何%ダウンロードされたか表示される。100%ダウンロードできたら初期画面が表示される(図 2)。

3.2 初期画面の構成

初期画面はデフォルトでは行に「都道府県」，列に「年」そしてデータ部分には「合計(DATA)」が入っている。画面の一番上の枠内には使用可能な変数リストが表示されている。使用可能な変数リストの中でデータを抽出する(たとえばデータ型は「収穫量」市町村は「福知山市」)。行列に表示させたい変数を変数リストより移動する(図 3)。(ただし変数リスト中の「DATA」は動かさない。左上の DATA と合計（整数）も通常は触れない。ただし，数値が小数の場合は「合計(整数)」を「合計(小数)」に変えたり,平均値が必要なら「平均」,割合が必要なら割合に設定することができる。

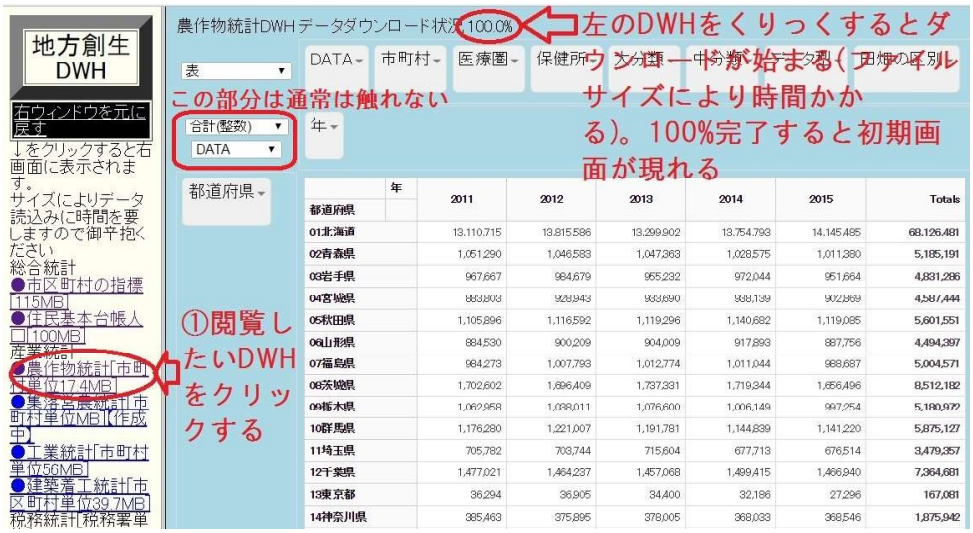

図 2 データダウンロード完了時の画面

図 3 使用できる変数リスト，行見出し，列見出しの関係

3.3 変数のドラグ＆ドロップ

ウェブ版 DWH の特色は,たとえ PC に Excel がなくてもウェブ上で,Excel ピボットテーブルのように変数をドラグ＆ドロップして自在に表示させることができる点にある。しかしあまり多くの変数を行列の見出しにいれると見にくくなるため,見出しにいれる変数の数は必要最小限にとどめる。そのためにはまず不必要な変数を上の変数リストに戻す。都道府県は不要なので戻す(図 4)。

図 4 行見出しから変数リストへの変数のドラグ＆ドロップ

3.4 データ型の選択

DWH では変数を分かりやすくするため,あらゆるデータを通じて一定のルールに従って命名している。たとえば,作付面積は ha(ヘクタール)で示され，収穫量や出荷量は t(トン)で示される異なるデータの型なので「データ型」と命名してある。なお「性別」というデータ型には「男」「女」という項目が含まれるので,男,女は「データ項目」と命名される。今回必要な「収穫量」は変数リスト中の「データ型」に入っている。右の▼をダブルクリックしてプルダウンメニューを出し，一旦 Select None をクリックして全てのチェックを外した後で「収穫量」のみをチェックして OK をクリックする(図 5)。

3.5 選択した変数のドラグ＆ドロップ

目標とするデータの分類(この場合は農作物)は「大分類→中分類→小分類→細分類」になっており，必ず,この順に選択を狭めてゆく。また小さな分類は必ず大きな分類の下に置く。まず「大分類」を変数リストより行にドラグ＆ドロップする(図 6)。

3.6 ドリルダウン

中分類を大分類の下にドラグ＆ドロップする(必ず小分類は大分類の下に配置する)。このように大きな分類から小さな分類に細かく表示させることをドリルダウンと呼ぶ(図 7)。

3.7 市町村の選択

市町村より京都府福知山市を選択する。市町村は 1700 以上もあるので，Filter に市町村名を入力することで容易に検索できる(図 8)。見つけたらチェックして OK ボタンを押す。

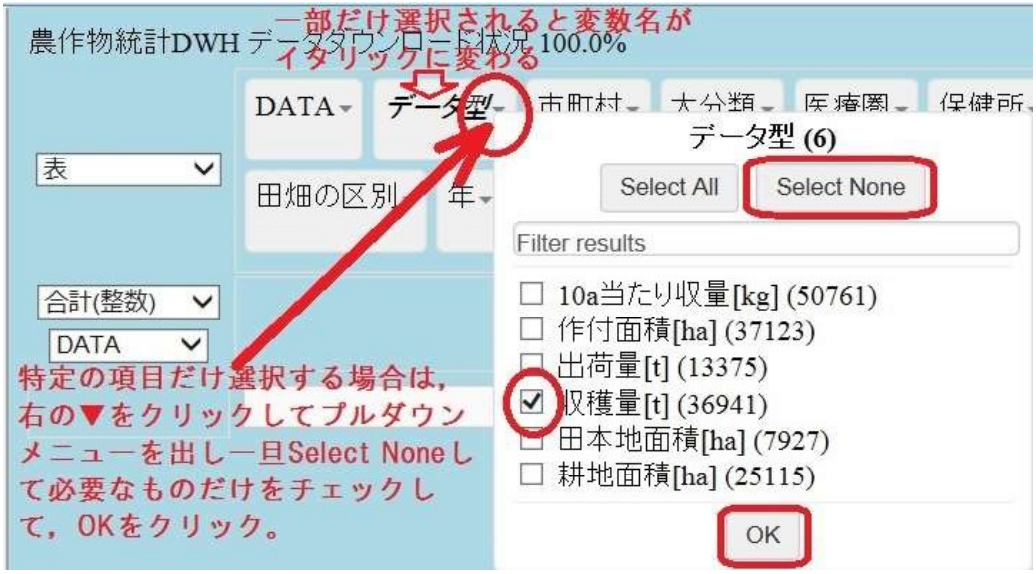

図 5 プルダウンメニューからのデータ型の選択

図 6 変数リストから行見出しへの変数のドラグ＆ドロップ

図 7 大分類から中分類へのドリルダウン

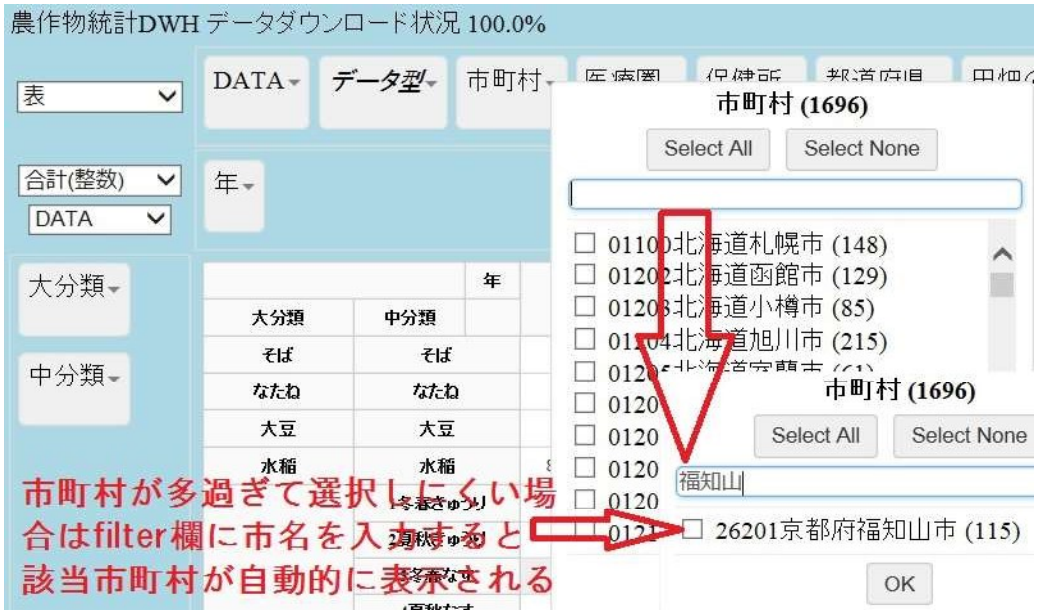

図 8 検索機能を用いた変数名の選択

3.8 完成

こうすることで福知山市の過去５年間における農作物の収穫量が農作物別に表示される(図 9)。(なお DWH ではメモリ節約のため数値が無かったりゼロのデータは略してある。よって以下に表示されていない農作物は福知山市では収穫されていないことを意味する。

大分類	中分類	年	2011	2012	2013	2014	2015	Totals
そば	そば		7	21	12	6	12	58
大豆	大豆		54	54	34	46	45	233
水稲	水稲		8,590	8,890	8,840	8,210	8,210	42,740
野菜(果菜類)	2夏秋きゅうり		277	356	257	278		1,168
麦類	1小麦		55	77		80	74	286
		Totals	8,983	9,398	9,143	8,620	8,341	44,485

図 9 福知山市５年間の農作物別収穫量の表示が完成した状態

3.9 棒グラフ

DWH には,棒グラフやヒートマップ表示機能もある。左上のウィンドウをプルダウンし,「表」から「バーチャート」「ヒートマップ」に変えることによって表示される(図 10)。

3.10 ヒートマップ

DWH はヒートマップを表示することも可能で,全体，行，列とそれぞれに対する割合の 3 種類を表示させることができる(図 11)。

3.11 ソート

DWH では,選択した行又は列によって全体をソートする機能がある。たとえば福知山市の水稲の収穫量が最も多かった年を知りたければ,水稲のセルをダブルクリックすれば昇順もしくは降順にソートできる(図 12)。また、ソートの向きは変数見出しの横に矢印で表示される。

図 10 タイトル

図 11 図 9 に棒グラフ表示を追加した結果

図 12　図 9 を水稲の収穫量でソートした結果

4.実際の活用例・・・北近畿の税務署管内別成人一人当たり酒の消費量

どの地域でどの酒がどれだけ消費されているのか?

酒には酒税という間接税がかかるので,国税庁の税務統計から知ることができる。しかし比較には一人当たりに換算することが必要であり,さらに酒の場合,未成年者を除いた成人一人当たりで比較するのが妥当であろう。税務統計には,年齢階級別人口は含まれていないので,人口統計と同一市町村,同一年で突合する必要がある。市町村別の毎年の性・年齢階級別人口としては住民基本台帳人口が適切であるが,国税庁が公表する統計は市町村単位ではなく税務署単位である。そうなると,各税務署の管轄に含まれる市町村を抽出し,さらに 20 歳以上人口を抽出し,さらに毎年の人口を入手しなければならず,手間のかかる作業となる。

こうした作業が DWH を用いると極めて容易になる。

4.1 北近畿各税務署別の清酒消費量

税務統計 DWH【酒税】より税務署別の酒消費量をだす。「税務署」を行見出しにドラグ＆ドロップし,北近畿を管轄する,福知山,舞鶴,宮津,峰山(以上,京都府)と,豊岡,和田山,柏原(以上,兵庫県)を選択する。さらに「酒」をダブルクリックして「清酒」を選択する。以下表示されるのは消費量(kL)である(図 13)。

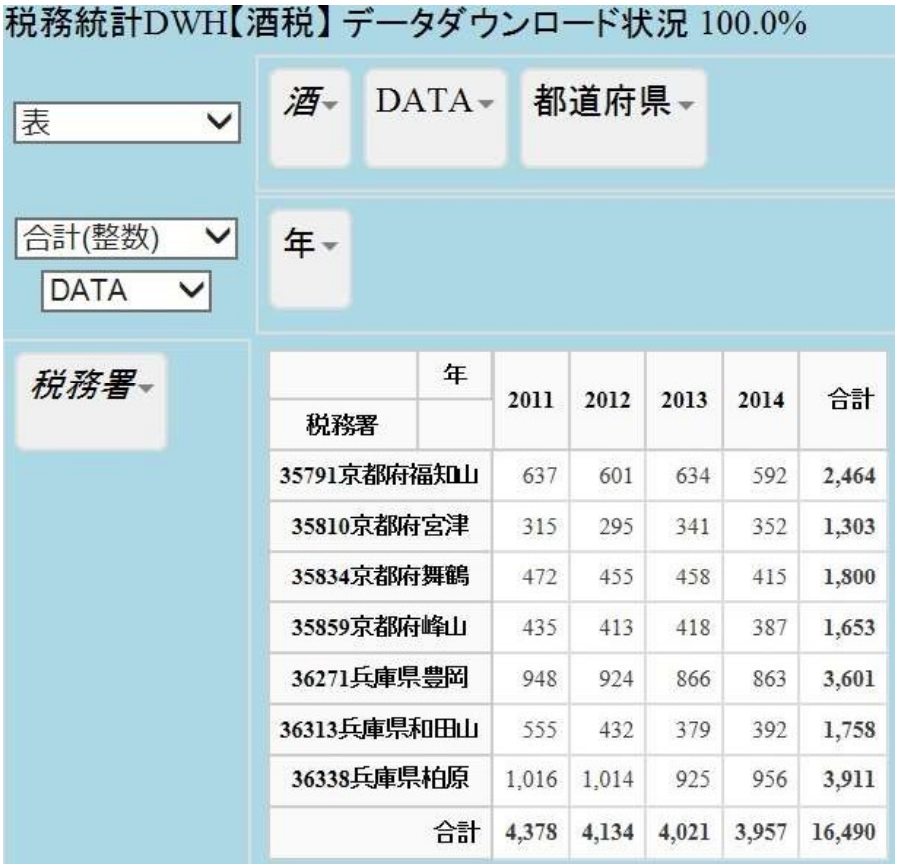

図 13　税務統計より北近畿の税務署別酒消費量を表示した状態

4.2 住民基本台帳人口からの成人人口の抽出

成人一人当たり消費量を算出するには，20 歳以上の成人人口を住民基本台帳人口 DWH より出して分母としなければならない。しかし住民基本台帳人口は市町村単位であって税務署単位ではない。税務署は，複数市町村を管轄するところが多い(たとえば福知山税務署は福知山市に加えて綾部市も管轄する(逆にひとつの自治体に複数の税務署があるところもあり，たとえば東京都世田谷区は区内に税務署が 3 つもある)。住民基本台帳人口 DWH は，市区町村単位の人口を，医療圏別，保健所並びに税務署管轄区域別に，かつ性・年齢階級別に自在に集計可能になっている。「年齢階級」より 20 歳未満のチェックをはずし「税務署」より酒税統計で抽出したものと同一の税務署を抽出する(図 14)。

税務署 \ 年	2011	2012	2013	2014	合計
35791京都府福知山	95,740	95,417	94,817	94,727	380,701
35810京都府宮津	39,216	38,839	38,417	38,179	154,651
35834京都府舞鶴	71,584	70,802	70,029	69,968	282,383
35859京都府峰山	49,636	49,193	48,663	48,511	196,003
36271兵庫県豊岡	102,391	101,468	100,624	100,160	404,643
36313兵庫県和田山	50,120	49,520	48,888	48,489	197,017
36338兵庫県柏原	92,694	92,196	91,675	91,225	367,790
合計	501,381	497,435	493,113	491,259	1,983,188

図 14 住民基本台帳より北近畿の税務署別成人人口を表示した状態

4.3 Excel への貼り付け→算出

DWH の表を Excel に貼り付け，清酒消費量を分子，成人人口を分母として一人当たり消費量を算出する。表 1 は算出された一人当たり消費量であり、Excel の「条件付き書式」によりグラフ化されている。北近畿の中でも福知山や舞鶴といった京都北部よりも、柏原等の兵庫県北部の方が清酒消費量が多いことがわかる。

4.4 酒種類別の地域特性

1 年ごとに酒の種類別に分析したではデータのぶれが大きいため,地域特性を知るためには 4 年間分を合計し,酒種別に分析することが有効であろう。2011～14 年間を合計し,酒の種類別にみたものが表 2 である。この表から、酒の種類別消費量の地域特性がうかがえる。兵庫県柏原税務署管内(丹波市及び篠山市)は清酒の消費量が多く,成人一人当たり 8.7 リットルと,舞鶴署管内の 5.1 リットルより高い。逆にビールは柏原署管内では少なく(16 リットル)に対して,舞鶴署管内は 22.4 リットルとなっている。

【表1】北近畿地域税務署別清酒成人一人当たり消費量

	2011	2012	2013	2014
清酒消費量(kL, 国税庁酒税統計)				
35791京都府福知山	637	601	634	592
35810京都府宮津	315	295	341	352
35834京都府舞鶴	472	455	458	415
35859京都府峰山	435	413	418	387
36271兵庫県豊岡	948	924	866	863
36313兵庫県和田山	555	432	379	392
36338兵庫県柏原	1,016	1,014	925	956
成人人口(住民基本台帳人口)				
35791京都府福知山	95,740	95,417	94,817	94,727
35810京都府宮津	39,216	38,839	38,417	38,179
35834京都府舞鶴	71,584	70,802	70,029	69,968
35859京都府峰山	49,636	49,193	48,663	48,511
36271兵庫県豊岡	102,391	101,468	100,624	100,160
36313兵庫県和田山	50,120	49,520	48,888	48,489
36338兵庫県柏原	92,694	92,196	91,675	91,225
成人一人当たり清酒消費量(L/人)				
35791京都府福知山	6.7	6.3	6.7	6.2
35810京都府宮津	8.0	7.6	8.9	9.2
35834京都府舞鶴	6.6	6.4	6.5	5.9
35859京都府峰山	8.8	8.4	8.6	8.0
36271兵庫県豊岡	9.3	9.1	8.6	8.6
36313兵庫県和田山	11.1	8.7	7.8	8.1
36338兵庫県柏原	11.0	11.0	10.1	10.5

【表2】北近畿税務署区域別成人一人当たり酒種別消費量(2011～14年平均)

税務署	ウイスキー	ビール	ブランデー	リキュール	単式蒸留しょうちゅう	原料用アルコール・スピリッツ	合成清酒	果実酒	清酒	甘味果実酒	発泡酒	連続式蒸留しょうちゅう	合計
35791京都府福知山	0.5	19.7	0.1	17.3	4.2	2.2	0.4	1.1	5.3	0.0	8.1	1.9	61.0
35810京都府宮津	0.4	18.6	0.0	11.1	3.4	1.5	0.4	1.6	7.0	0.0	7.0	1.7	52.8
35834京都府舞鶴	0.8	22.4	0.1	14.6	5.2	2.2	0.3	1.8	5.1	0.1	9.3	1.7	63.7
35859京都府峰山	0.3	20.9	0.0	9.7	4.3	1.4	0.5	0.9	6.9	0.1	8.4	1.5	55.1
36271兵庫県豊岡	0.4	25.6	0.0	15.2	4.2	2.3	0.6	0.9	7.3	0.0	7.9	1.7	66.1
36313兵庫県和田山	0.4	17.7	0.0	14.7	4.3	1.9	0.6	1.0	7.4	0.0	6.5	2.0	56.5
36338兵庫県柏原	0.3	16.0	0.0	13.1	3.2	1.8	0.2	0.7	8.7	0.0	5.5	1.6	51.2

5.まとめ

市町村別の詳細な地域データは以前から公表されていたが，量が膨大であるため刊行物には収録されず，たとえば厚生労働省の統計情報部において分厚い集計表(なかには連続紙のコンピュータープリントアウトそのまま，というものもある)を閲覧させてもらう，いわゆる「閲覧公表」というかたちでしか提供されなかった。それが e-STAT 上で Excel や csv ファイルで提供されるようになり，従来は困難だった市町村別データが容易に入手できるようになった。著者(岡本)が厚生労働科学研究として DWH 化ととりくんだのは 2010 年にさかのぼるが，そろそろ 5～6 年分の市町村別データが Excel や csv ファイルで入手可能となってきた。そうなると経時的な推移を知りたくなるが，e-STAT 上で公表されるファイルは各年分のみであって，長期間のデータを縦覧できるようにはなっていない。政府統計は統計法によって収集，集計そして公表されており，各年に承認された形式以外に，たとえば「過去 5 回分の国勢調査を市町村別に経年推移をおえるように加工する」ことは特別な手続きをふまなければ行なえない。ユーザーは，もし経年推移を知りたければ e-STAT から 5 回分のデータをダウンロードして自らカット&ペーストして加工しなければならない。

ならいっそ，複数年にわたる市町村別データをあらかじめ加工すればいい，が本プロジェクトの着想である。しかし，その作業は決して容易ではない。市町村名の表記の「揺れ」はもとより，今世紀に入ってからの「平成の大合併」により,かつて 3000 以上あった市町村は 1700 余りに激減した。表記の統一，消滅した市町村が現在のどの市町村に該当するか，を結合するマスターファイルが必要だった。また，膨大なファイルをダウンロードする API プログラム，クロス表をキューブ化する逆ピボットテーブル手法等も不可欠な技術である。

何よりのチャレンジは，ウェブ上でピボットテーブルを実現する Javascript プログラムである。現在のプログラムはサーバーからいったん全データをダウンロードし，クライエントのコンピューター内で処理する仕組みであるため，ダウンロードに時間がかかる。これらデータ処理をサーバー内で処理し，結果のみを返すシステムの構築が次なる課題である。クロス表をキューブ化する処理によりデータ量は，100 行×100 列のクロス表なら 1 万セルだが，1000 行×1000 列では 100 万セル，と指数関数的に膨張するので現在のシステムでは早晩行き詰まる。また成人一人当たり消費量を算出するために,分子の酒消費量と分母の成人人口を二つの DWH から抽出した後,Excel にカット&ペーストして算出しなければならなかった。理想的には「分子に清酒消費量」「分母に成人人口」と指定するだけで,指標が自動的に算出されるようなシステムが望まれる。さらには,まったく異なる指標と指標との相関を市町村単位で総当たり的に分析することによって,未知の関係を明らかにするデータマイニング的な活用も今後の課題である。億単位のデータを短時間で処理できる Hadoop のような並列処理サーバーを用いれば，今回は手作業で行った異なるデータ間の突合も自動的に行えるであろう。

税務統計と住民基本台帳人口といった異なる統計データを,たとえば同一市町村で結合して,成人一人当たり酒の消費量といった正確な指標により地域特性を明らかにできることを示した。しかしなが

ら,国税庁の出す税務統計と総務省の出す住民基本台帳統計,さらには厚生労働省が出す人口統計とは,体裁や表記が微妙に異なっており(たとえば龍ケ崎と竜ヶ崎)その結合は容易ではない。地域データも,市町村,市区町村(指定都市は区単位),保健所管轄区域,医療圏さらには税務署単位というぐあいにその区分は統一されていない。DWH に加工することにより,税務署単位でも,保健所単位でも,はたまた医療圏単位でも自在に集計できるようになる。たとえば健康増進を目的に一人当たり飲酒量と人口当たり肝疾患有病率の相関を分析したり，居酒屋の出店計画のためにビールより清酒が好まれる地域はどこか,といった分析も容易になり，地方創生のための有効なツールとして期待される。

6. 謝辞

本研究は内閣府地方活性化加速化交付金「京都府北部地域連携都市圏　地（知）の拠点推進事業」の助成を受けた。

<<参考文献>>

(1)　厚生労働科学研究健康安全・危機管理対策総合研究事業「保健医療福祉計画策定のためのデータウェアハウス構築に関する研究」 (研究代表者:岡本悦司)平成 27 年度　総括・分担研究報告書(2016 年 3 月)

(2)　厚生労働科学統計情報総合研究事業「OLAP(多次元データベース)による医療統計の公表手法開発に関する研究」(研究代表者:岡本悦司)平成 22・23 年度　総合研究報告書(2012 年 3 月)