福知山公立大学
The University of Fukuchiyama

地方創生加速化交付金プロジェクト

地方創生データウェアハウス構築に関する研究

成果報告書

岡本悦司，神谷達夫

福知山公立大学
地域経営学部医療福祉マネジメント学科
データウェアハウス構築チーム
2017 年 3 月

成果要旨

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ウェブ上で公開(japanreview.com)した。その概要と使用法を紹介するとともに今後の展開の方向を示す。

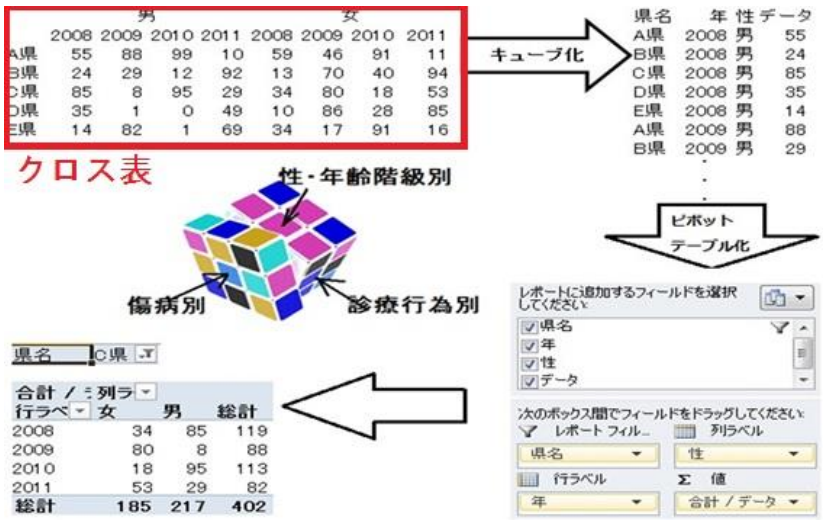
1. データウェアハウスとは

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。

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，鯵ケ沢町と鰺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。



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

収録済データは現時点では以下の通りであり，順次拡大してゆく。

総合統計
●国勢調査(2015 年)
●市区町村のすがた
●住民基本台帳人口
●将来推計人口
●外国人人口
産業統計
●農作物統計
●集落営農統計【2015 年度のみ】
●工業統計
●建築着工統計[市区町村単位]
税務統計
__国税[税務署単位]（データ源国税庁）
●源泉所得税
●申告所得税
●酒税
__住民税[市町村単位](データ源総務省)
●市町村民税課税状況【主たる所得別】
●住民税所得割課税状況【所得種類別】
●軽自動車税

3. DWH の使用法

DWH の使用法を農作物統計より「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(農作物統計)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。

3.2 初期画面の構成

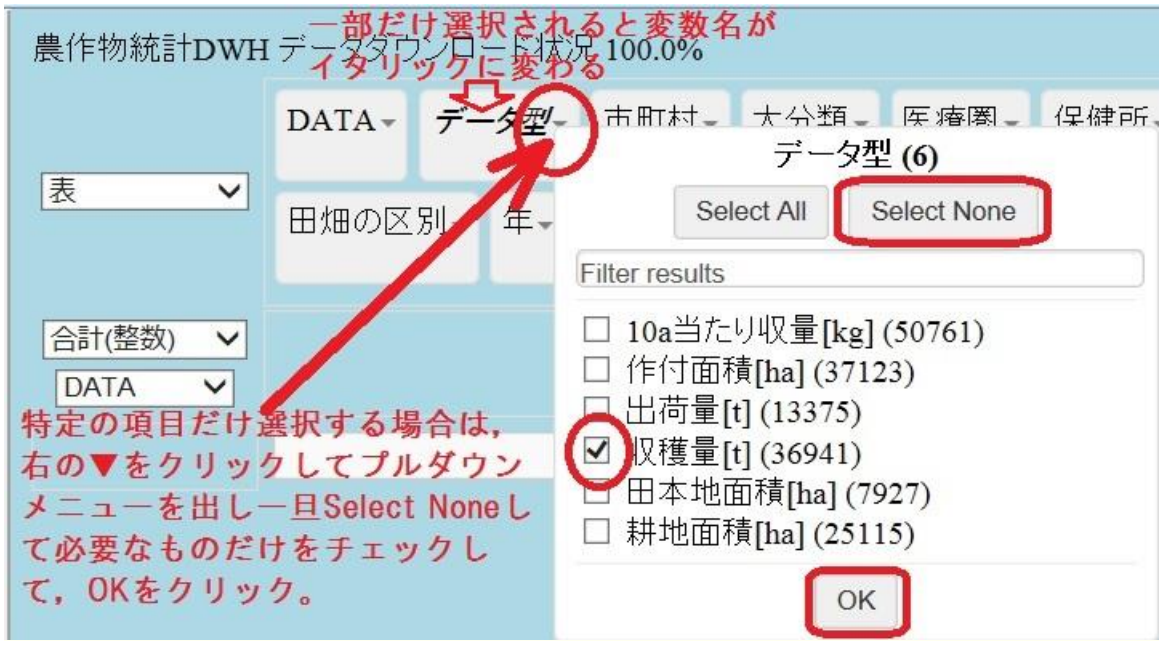
初期画面はデフォルトでは行に「都道府県」，列に「年」そしてデータ部分には「合計(DATA)」が入っている。画面の一番上の枠内には使用可能な変数リストが表示されている。使用可能な変数リストの中でデータを抽出する(たとえばデータ型は「収穫量」市町村は「福知山市」)。行列に表示させたい変数を変数リストより移動する***(ただし変数リスト中の「DATA」は動かさない)***。左上の DATA と合計（整数）も通常は触れない(ただし，数値が小数の場合は「合計(整数)」を「合計(小数)」に変えたり，平均値が必要なら「平均」，割合が必要なら割合)。

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。

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。



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

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

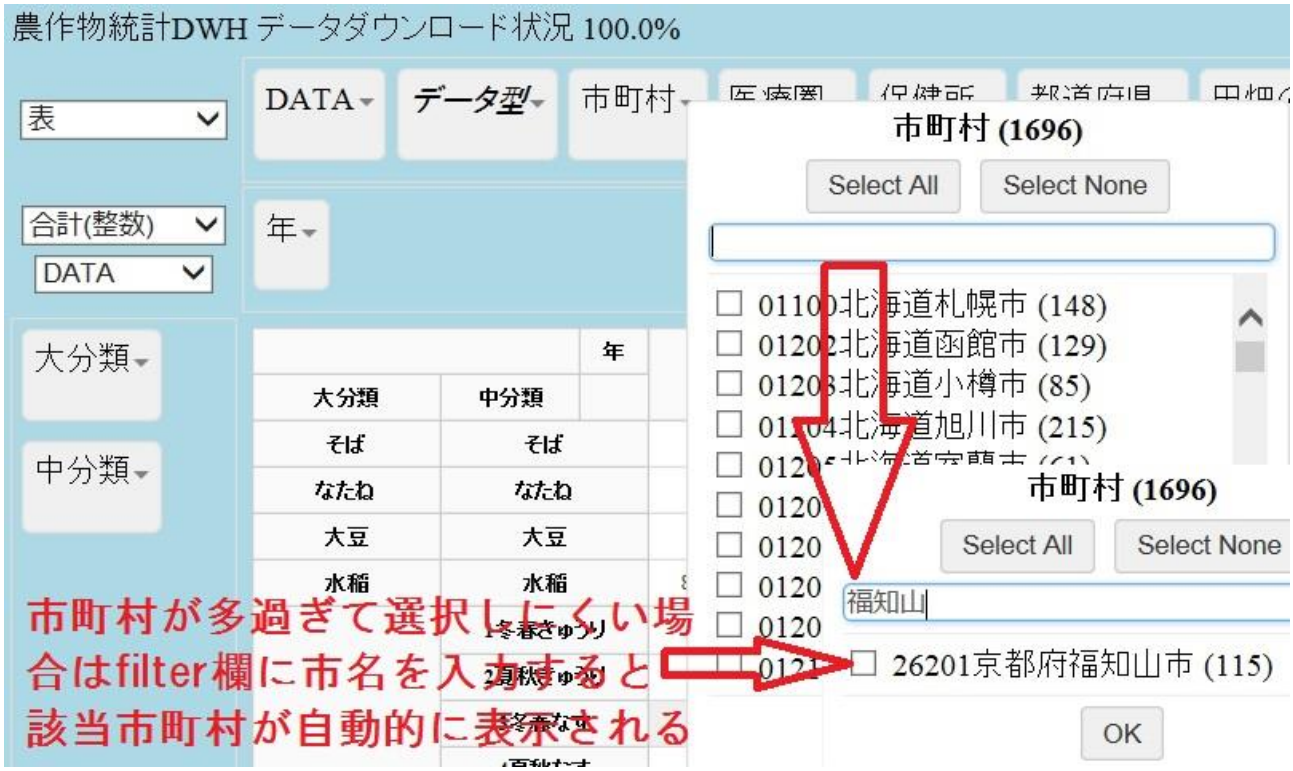
3.6 ドリルダウン

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。

大分類	中分類 \ 年	2011	2012	2013	2014	2015	Totals
そば	そば	31,178	43,780	32,698	30,367	33,711	171,734
なたね	なたね	1,530	1,462	1,231	1,179	2,141	7,543
大豆	大豆	218,393	235,679	199,302	231,149	242,389	1,126,912
水稲	水稲	8,395,691	8,519,099	8,602,965	8,435,537	7,985,894	41,939,186
野菜(果菜類)	1冬春きゅうり	234,953	226,353	230,459	213,818	212,782	1,118,365
	2夏秋きゅうり	138,779	146,215	131,197	128,390	134,554	679,135
	3冬春なす	96,214	94,027	98,557	102,223	95,970	486,991
	4夏秋なす	50,043	53,276	50,725	50,996	51,937	256,977
	5冬春トマト	246,673	238,218	264,008	257,023	250,234	1,256,156
	6夏秋トマト	191,907	209,555	204,585	205,798	204,172	1,016,017

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。みつけたらチェックして OK。



3.8 完成

こうすることで福知山市の過去５年間の農作物の収穫量が農作物別に表示される（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	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

3.9 棒グラフ

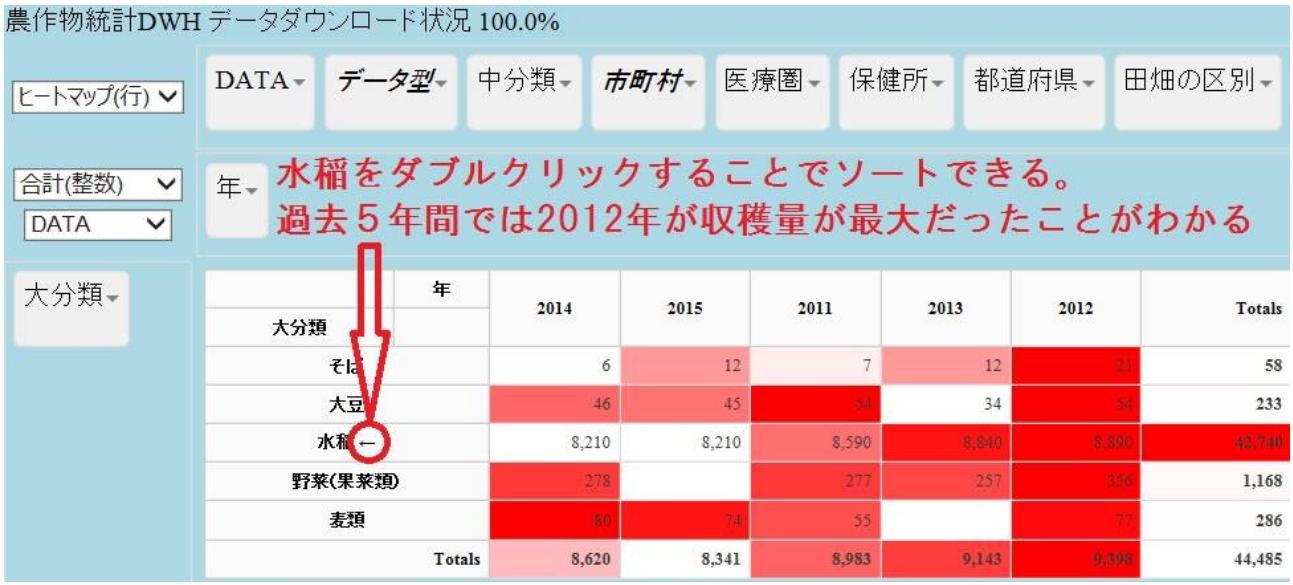
DWH には,棒グラフやヒートマップ表示機能もある。左上のウィンドウをプルダウンし,「表」から「バーチャート」「ヒートマップ」に変えることによって表示される

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。

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(向きは変数見出しの横に矢印で表示される)。



4. 今後の展開

市(区)町村単位で公開されている多数の統計調査データを加工してデータウェアハウスとし，かつウェブ上で公開して誰でも利用可能とした。

キューブ化されたデータは csv 形式でサーバーに入っているが，現在の JavaScript プログラムは，選択した csv データを一旦全てユーザーの PC にダウンロードすることが必要であり，ファイルが大きくなるとダウンロードに時間がかかる。また一つの統計データごとにダウンロードしなければならないので，たとえば国勢調査と税務統計とを同時に表示することはできない。データ量も 100 行×100 列のクロス表はキューブ化しても 1 万行であるが, 1000 行×1000 列では 100 万行と指数関数的に大きくなる。

たとえば，経済センサスは市区町村ではなく，大字単位のデータが収録されている。市町村は約 1700 だが，大字は約 17 万あり，もし大字単位のデータウェアハウスを構築するためには，データ量は 100 倍にもなる。そうなるとユーザーの PC に一旦ダウンロードすることを必要とする現在のシステムでは限界があり，やはりサーバー側で，膨大なデータより迅速に抽出，集計できるシステムが望まれる。現在，e-STAT 等で公開されているクロス表は膨大であり，それらを全てキューブ化すると億ないしはそれ以上の行数となると予想される。

かかる「ビッグデータ」の処理のため，多数の CPU をつないで並列処理できる HADOOP サーバーを構築した。下写真のように 24CPU を並列稼働させることにより，億単位の件数であっても短時間で処理できるようになった。データウェアハウスへの応用は年度内に間に合わなかったが，データウェアハウスを HADOOP による並列処理できるシステムに向上させ，可能な限り早くウェブ上での公開に発展させる方針である。