Investigating the Relationship between **Social Media Content** and **Real-time Observations** for Urban Air Quality and Public Health

Authors
Marina Riga and Kostas Karatzas

Marina Riga, PhD student
Informatics Systems and Applications Group (ISAG)
Department of Mechanical Engineering
Aristotle University of Thessaloniki (AUTh), Greece
Email: mriga@isag.meng.auth.gr
Web: http://isag.meng.auth.gr/~mriga

4 June 2014

Introduction

Web 2.0

The change

Participatory Sensing (PS)

People “sensing” the environment

Applications and Services

Health (pollen diary)

Environment (pollution of lakes)

Urban (social events, condition of roads and urban space)

Image from AppAppeal.com website

Question: Can they operate supplementary?

Diversity / Correlation of observations

… Our aim is to investigate the relation between these two heterogeneous sources of data.

Area of Interest

Urban Air Quality and Public Health

Information Systems

Current services: official monitoring stations giving actual concentrations (numerical)

PS services: sensitive / involved groups giving subjective estimations (textual)

Which will be the two heterogeneous sources to be utilized?

Twitter

+ Wide adoption from users / citizens

+ Rich source of information

+ Personal opinions / observations / reports

ECMWF (European Centre for Medium-Range Weather Forecasts)

+ Historical atmospheric data

Data from Twitter

Crawling

Keywords: air quality, atmosphere, pollution, air pollutants, medication, symptoms, allergies, pollen, sneezing, itching, ...

Time span: February to June 2013

Geo-location: mainly in Europe and UK

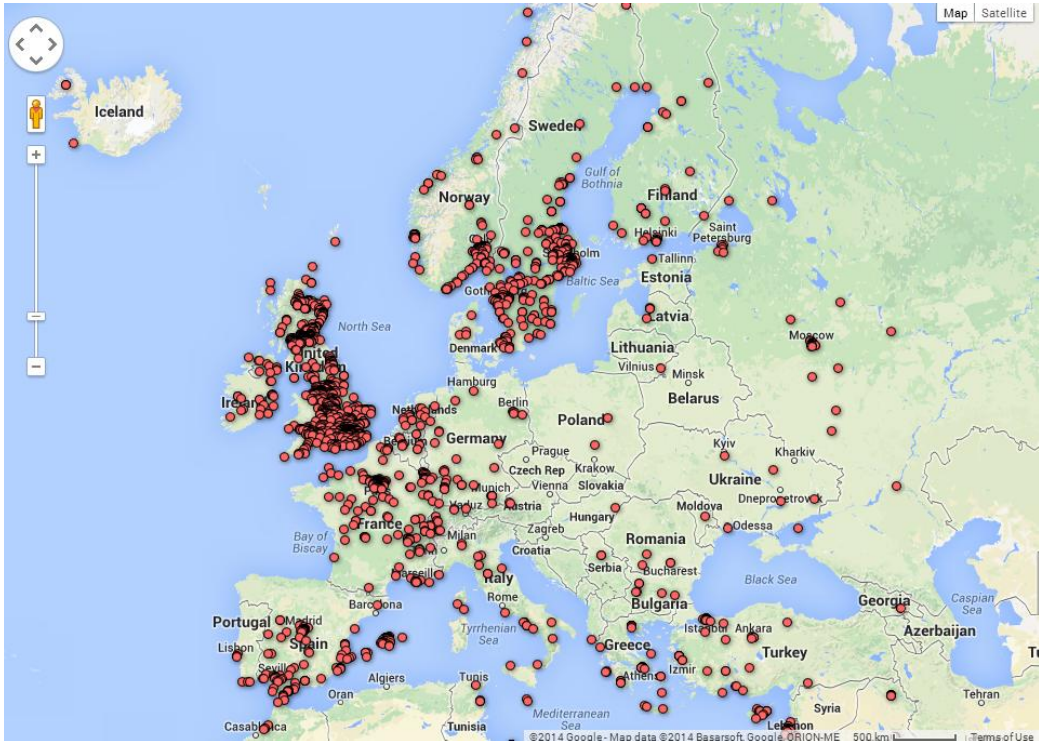
Preprocessing

Remove redundant content

- hyperlinks
- stop words
- usernames (@)
- hashtags (#)
- emoticons

Remove RTs

Result: 17,560 unique tweets



Data from ECMWF

Retrieving (batch request)

Parameters: wind speed, air temperature, skin temperature

- no available pollutants’ or pollen concentrations

One-by-one matching of tweets and official measurements

on the basis of **timestamp** and **geolocation**

How to combine the available data?

Heterogeneity

Textual and Numerical

The Feature Vector Model (in general): $d_i = [f_1, f_2, \dots, f_n]$

Represent text into a structured form

Bag of words (unigrams, n-grams)

But.. there is a need to:

Overcome the increased dimensionality of data

Include *not-so-frequent* words

We create a **bag of** **sets** **of words**

Based on the **most frequent** used words in the collection

Additional words attached to sets empirically

Taking into account issues of **polysemy**, **homonymy** and **semantic similarity**

Bag of sets of words

#	Words in set	Unified Concept
1	air, atmosphere, atmospheric	atmosphere
2	eyes, nose, throat, head, lungs, skin, heart, chest, body	body part / organ
3	pollution, pollute, pollutant(s)	pollution
4	itch, itching, itchiness	itch
5	sneeze, sneezing	sneeze
6	cough, coughing	cough
7	running, runny nose	runny
8	flu, sick, cold, ill, fever, disease, hay fever, asthma	medical condition
9	quality	quality
10	problem, difficulty	problem
11	allergy (ies), allergic, sensitive	allergy
12	food, eat	food
13	pollen	pollen
14	hospital, clinic, doctor	hospital
15	medication, medicine, pills	medication
16	car, vehicle, bus, bike, motor	vehicle
17	pets, dogs, cats, birds	pets
18	particles, particulates, PM, PM_{10}, $PM_{2.5}$, ozone, O_3	PM / O_3
19	hate, horrible, hell, crazy, killing, ugh	bad feelings
20	happy, funny, yeah	good feelings

Bag of sets of words

#	Words in set	Unified Concept
1	**air, atmosphere, atmospheric**	**atmosphere**
2	eyes, nose, throat, head, lungs, skin, heart, chest, body	body part / organ
3	pollution, pollute, pollutant(s)	pollution
4	itch, itching, itchiness	itch
5	sneeze, sneezing	sneeze
6	cough, coughing	cough
7	running, runny nose	runny
8	flu, sick, cold, ill, fever, disease, hay fever, asthma	medical condition
9	quality	quality
10	problem, difficulty	problem
11	allergy (ies), allergic, sensitive	allergy
12	food, eat	food
13	pollen	pollen
14	hospital, clinic, doctor	hospital
15	medication, medicine, pills	medication
16	car, vehicle, bus, bike, motor	vehicle
17	pets, dogs, cats, birds	pets
18	particles, particulates, PM, PM_{10}, $PM_{2.5}$, ozone, O_3	PM / O_3
19	hate, horrible, hell, crazy, killing, ugh	bad feelings
20	happy, funny, yeah	good feelings

Bag of sets of words

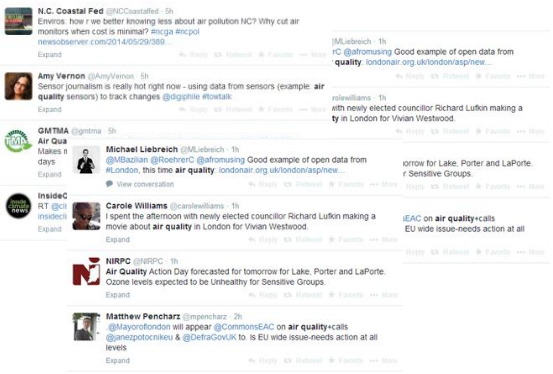
#	Words in set	Unified Concept
1	air, atmosphere, atmospheric	atmosphere
2	eyes, nose, throat, head, lungs, skin, heart, chest, body	body part / organ
3	pollution, pollute, pollutant(s)	pollution
4	itch, itching, itchiness	itch
5	sneeze, sneezing	sneeze
6	cough, coughing	cough
7	running, runny nose	runny
8	**flu, sick, cold, ill, fever, disease, hay fever, asthma**	**medical condition**
9	quality	quality
10	problem, difficulty	problem
11	allergy (ies), allergic, sensitive	allergy
12	food, eat	food
13	pollen	pollen
14	hospital, clinic, doctor	hospital
15	medication, medicine, pills	medication
16	car, vehicle, bus, bike, motor	vehicle
17	pets, dogs, cats, birds	pets
18	particles, particulates, PM, PM_{10}, $PM_{2.5}$, ozone, O_3	PM / O_3
19	hate, horrible, hell, crazy, killing, ugh	bad feelings
20	happy, funny, yeah	good feelings

Bag of sets of words

#	Words in set	Unified Concept
1	air, atmosphere, atmospheric	atmosphere
2	eyes, nose, throat, head, lungs, skin, heart, chest, body	body part / organ
3	pollution, pollute, pollutant(s)	pollution
4	itch, itching, itchiness	itch
5	sneeze, sneezing	sneeze
6	cough, coughing	cough
7	running, runny nose	runny
8	flu, sick, cold, ill, fever, disease, hay fever, asthma	medical condition
9	quality	quality
10	problem, difficulty	problem
11	allergy (ies), allergic, sensitive	allergy
12	food, eat	food
13	pollen	pollen
14	hospital, clinic, doctor	hospital
15	medication, medicine, pills	medication
16	car, vehicle, bus, bike, motor	vehicle
17	pets, dogs, cats, birds	pets
18	particles, particulates, PM, PM_{10}, $PM_{2.5}$, ozone, O_3	PM / O_3
19	**hate, horrible, hell, crazy, killing, ugh**	**bad feelings**
20	happy, funny, yeah	good feelings

Schematic representation of forming data

Moving from unstructured to structured data



N.C. Coastal Fed @NCCoastalfed · 5h
Enviros: how r we better knowing less about air pollution NC? Why cut air monitors when cost is minimal? #ncga #ncpol newsobserver.com/2014/05/29/389...
Expand

Amy Vernon @AmyVernon · 5h
Sensor journalism is really hot right now - using data from sensors (example: **air quality** sensors) to track changes @digiphile #towtalk
Expand

GMTMA @gmtma · 5h
Air Qua
Makes r
days
Expand

InsideC
RT @cli
insidecli
Expand

@MLiebreich · 1h
rC @afromusing Good example of open data from
quality: londonair.org.uk/london/asp/new...

rolewilliams · 1h
with newly elected councillor Richard Lufkin making a
ty in London for Vivian Westwood.

orrow for Lake, Porter and LaPorte.
r Sensitive Groups.

sEAC on **air quality**+calls
EU wide issue-needs action at all

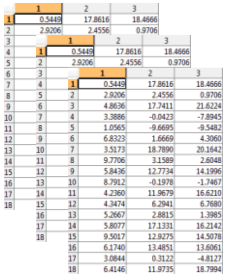
Michael Liebreich @MLiebreich · 1h
@MBazilian @RoehrerC @afromusing Good example of open data from #London, this time **air quality**: londonair.org.uk/london/asp/new...
View conversation

Carole Williams @carolewilliams · 1h
I spent the afternoon with newly elected councillor Richard Lufkin making a movie about **air quality** in London for Vivian Westwood.
Expand

NIRPC @NIRPC · 1h
Air Quality Action Day forecasted for tomorrow for Lake, Porter and LaPorte. Ozone levels expected to be Unhealthy for Sensitive Groups.
Expand

Matthew Pencharz @mpencharz · 2h
.@Mayoroflondon will appear @CommonsEAC on **air quality**+calls @janezpotocnikeu & @DefraGovUK to. Is EU wide issue-needs action at all levels
Expand

Reply
Retweet
Favorite
More



	1	2	3
1	0.5449	17.8616	18.4666
2	2.9206	2.4556	0.9706

	1	2	3
1	0.5449	17.8616	18.4666
2	2.9206	2.4556	0.9706

	1	2	3
1	0.5449	17.8616	18.4666
2	2.9206	2.4556	0.9706
3	4.8636	17.7411	21.6224
4	3.3886	-0.0423	-7.8945
5	1.0565	-9.6695	-9.5482
6	6.8323	1.6669	4.3060
7	3.5173	18.7890	20.1642
8	9.7706	3.1589	2.6048
9	5.8436	12.7734	14.1996
10	8.7912	-0.1978	-1.7467
11	4.2360	11.9679	16.6210
12	4.3474	6.2941	6.7680
13	5.2667	2.8815	1.3985
14	5.8077	17.1331	16.2142
15	9.5017	12.9275	14.5078
16	6.1740	13.4851	13.6061
17	3.0844	0.3122	-4.8127
18	6.4146	11.9735	18.7994

n features
(a bag of 20 sets of words + 3 numerical observations)

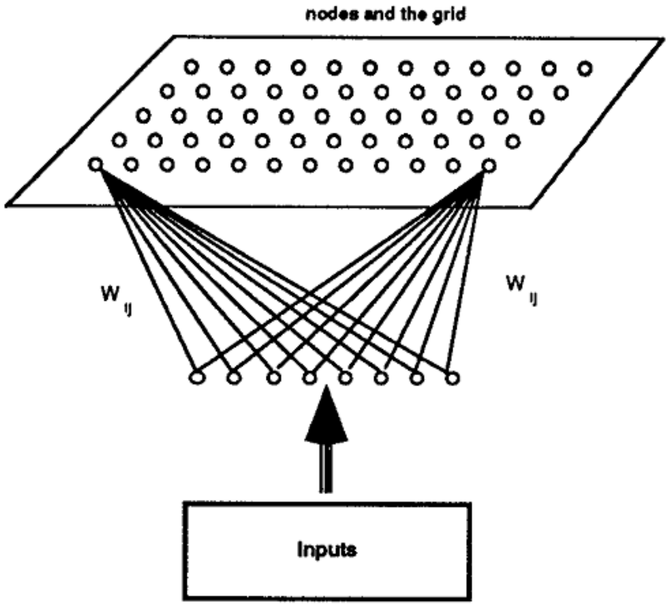
$$\overbrace{\begin{bmatrix} 0\;0\;1\;1\;1\;0\;0\;1 & \cdots & 0\;5.3\;13\;10 \\ 1\;0\;1\;0\;0\;0\;0\;0 & \cdots & 1\;0.4\;22\;25 \\ \vdots & \ddots & \vdots \\ 0\;1\;1\;0\;0\;0\;1\;0 & \cdots & 0\;9.1\;10\;7 \end{bmatrix}}^{}\left.\vphantom{\begin{bmatrix}0\\0\\0\\0\end{bmatrix}}\right\} \boldsymbol{m} \text{ tweets}$$

Self-Organizing Map (SOM)

Kohonen's Self – Organizing Maps (SOM)

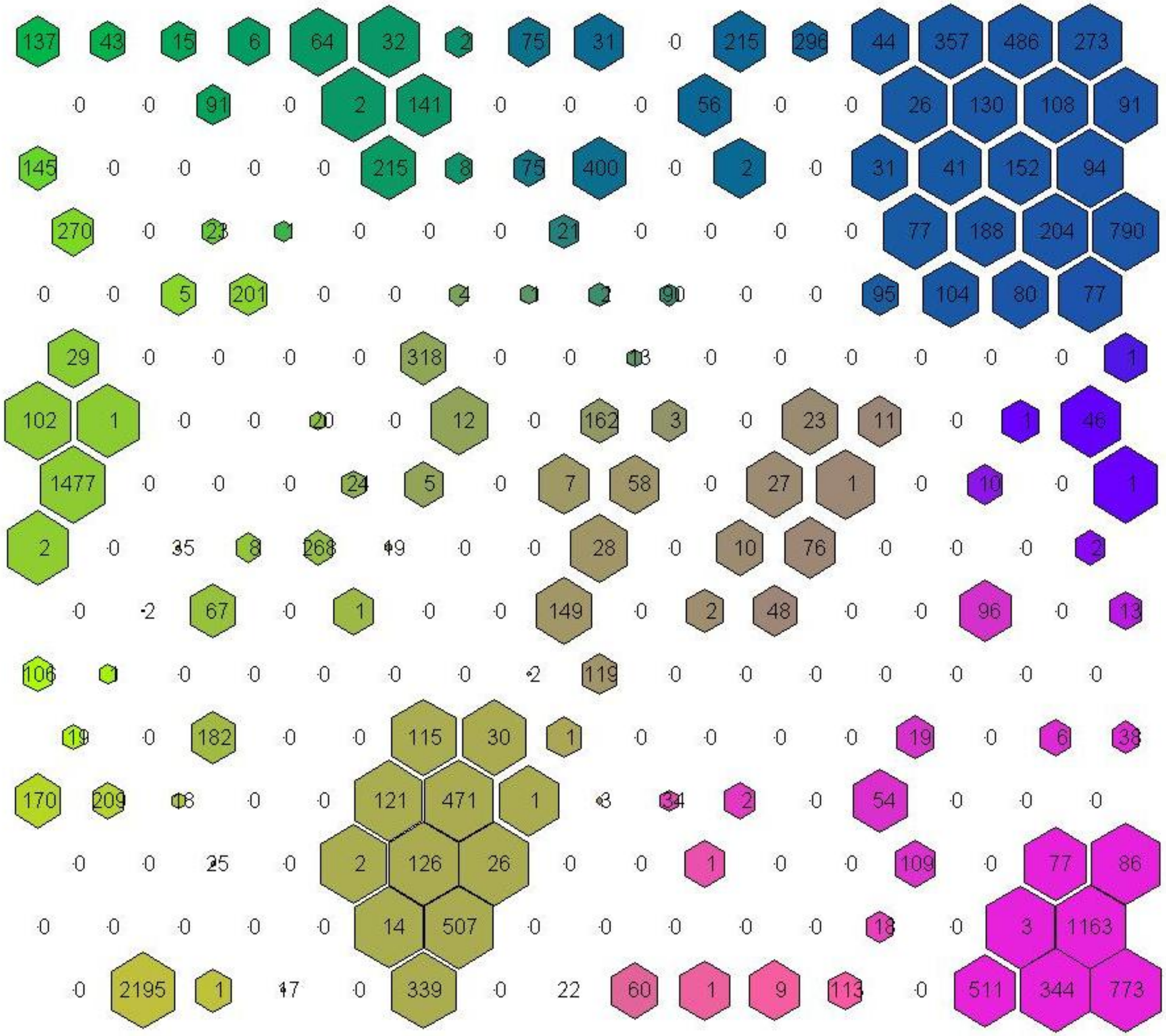
- Unsupervised learning method
- Maps high dimensional data into low (2D) dimensional space
- Preserves their spatial correlation
- **Similarity metric**: Euclidean Distance
- **Clusters**: with K-means

We feed the formed feature vectors as input to the SOM algorithm

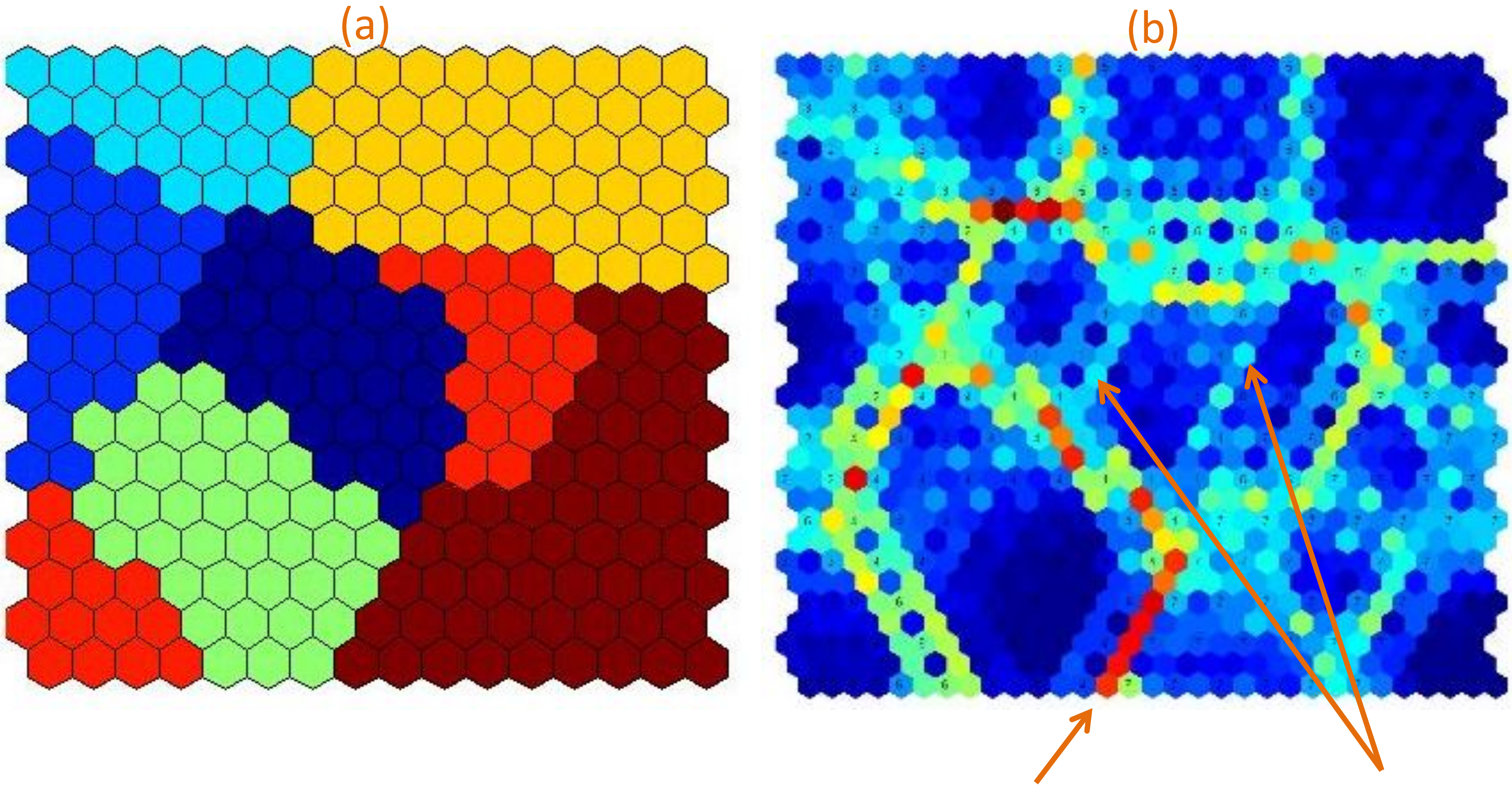


Kohonen's feature map

Results (1/3) – Hit map



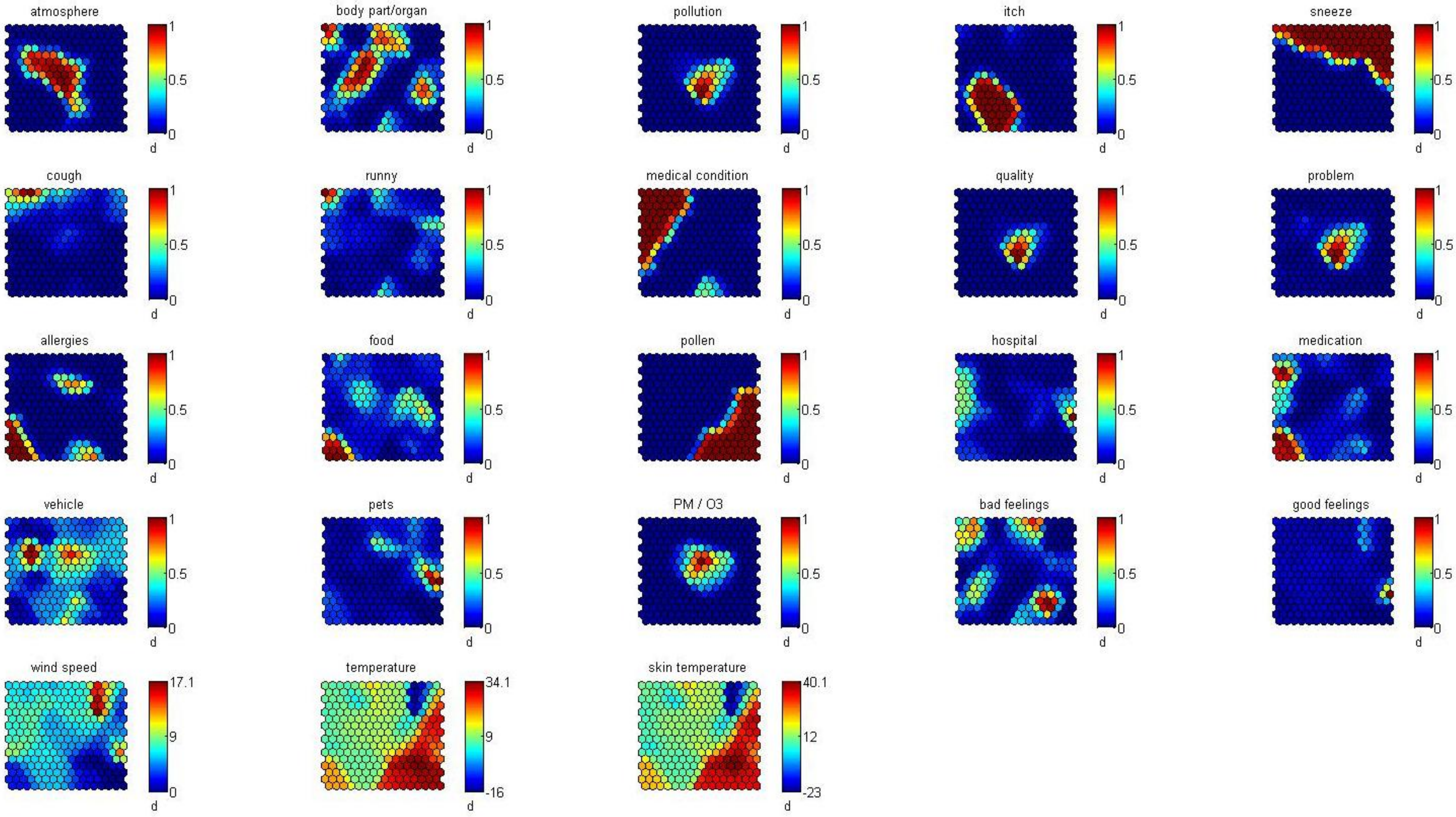
Results (2/3) – (a) Clusters and (b) U-matrix



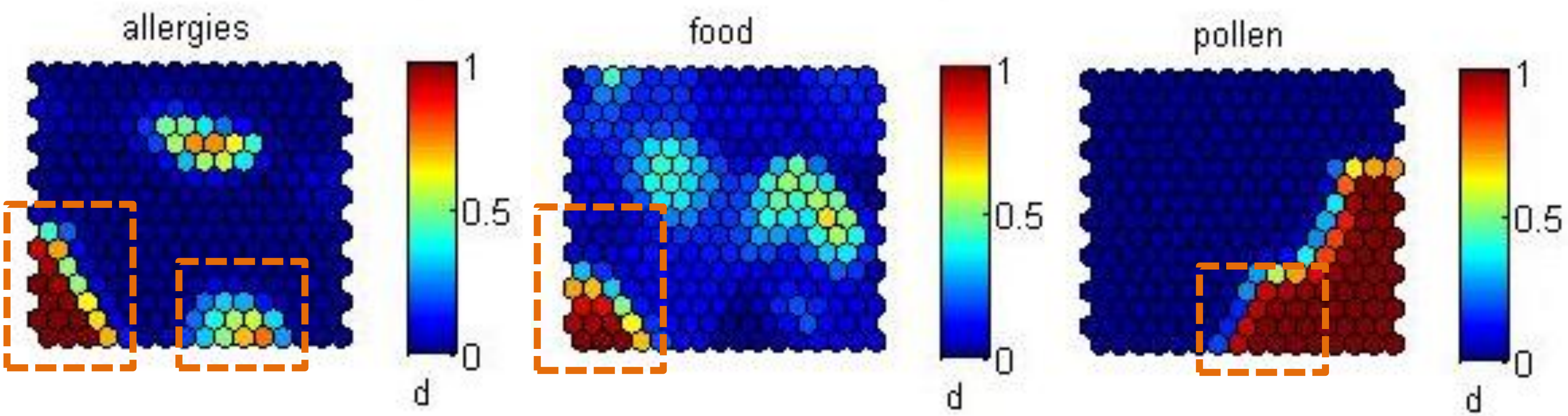
well defined boundaries

unclear boundaries

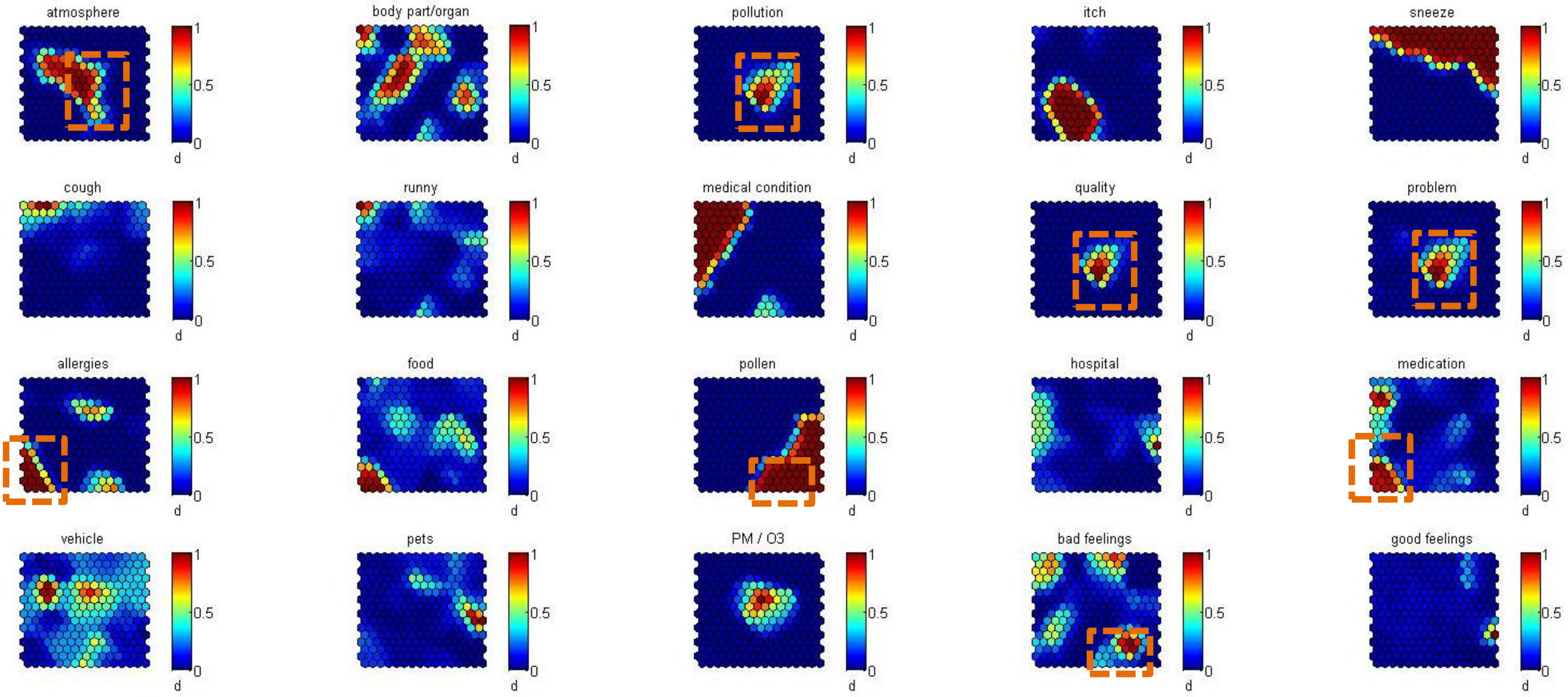
Results (3/3) – Component planes



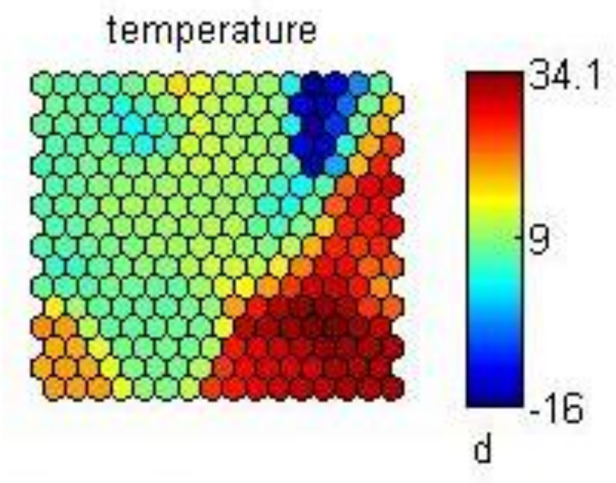
A. Relations between Sets of Words (Tweets)



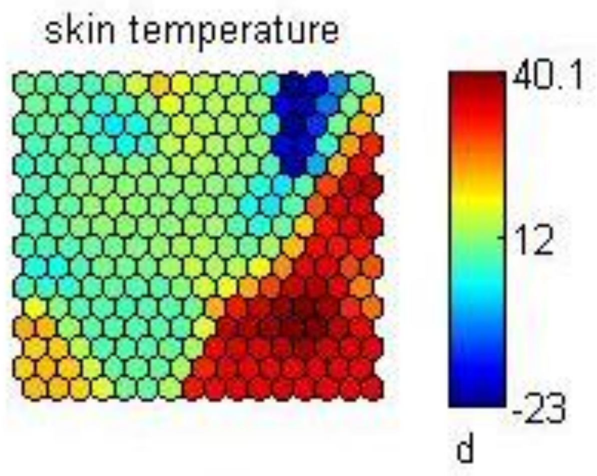
A. Relations between Sets of Words (Tweets)



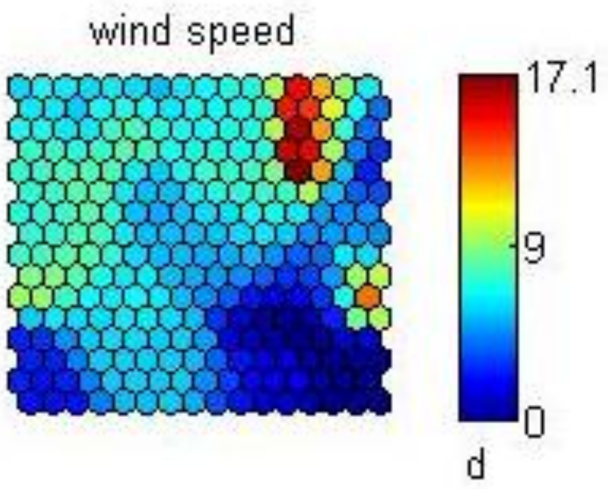
B. Relations between Official Observations (ECMWF)



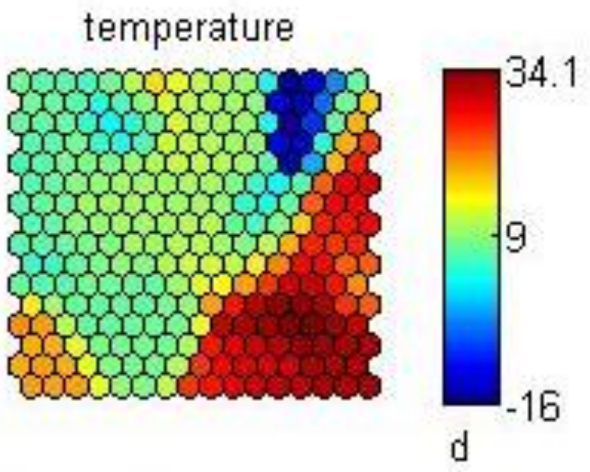
≈



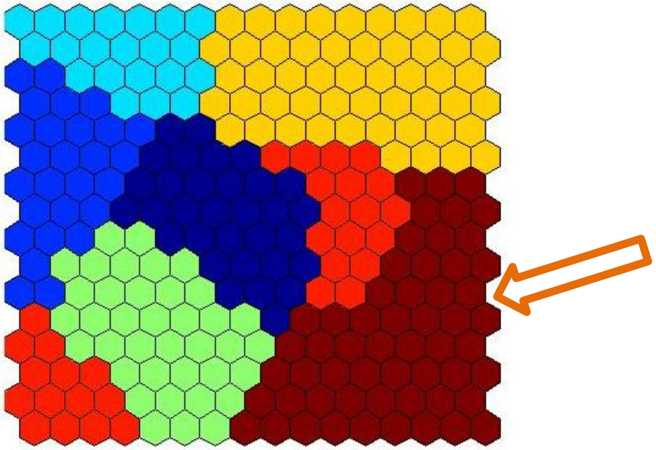
B. Relations between Official Observations (ECMWF)

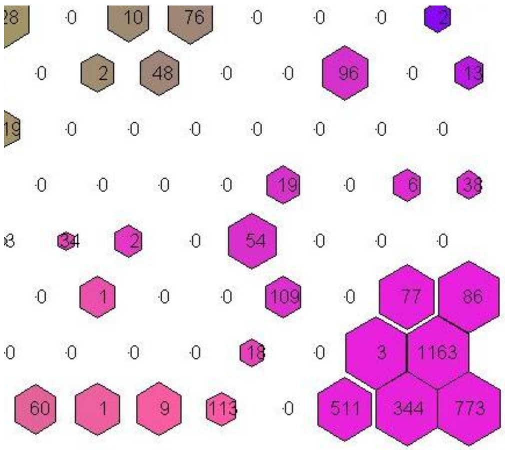


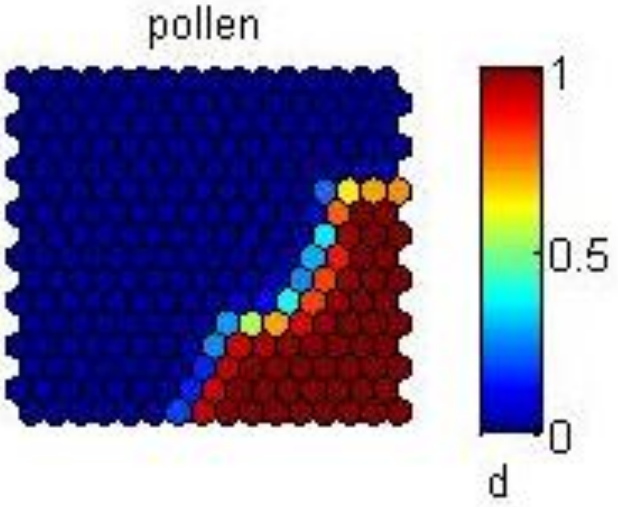
≠

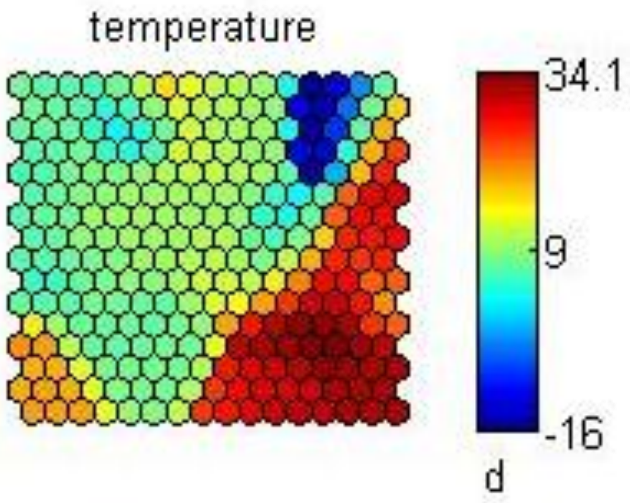


C. Relations of Sets of Words & Official Observations

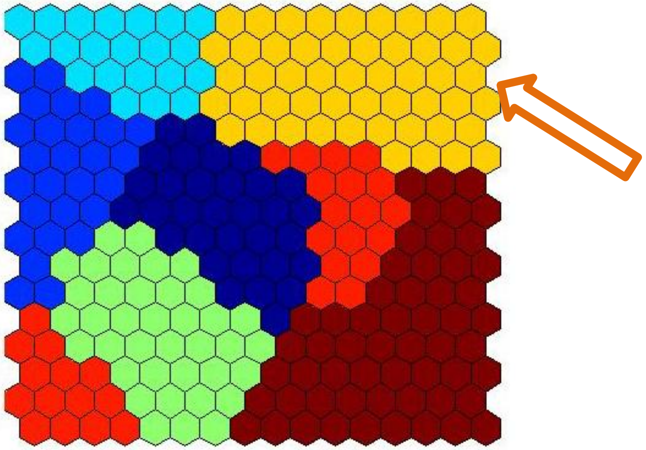


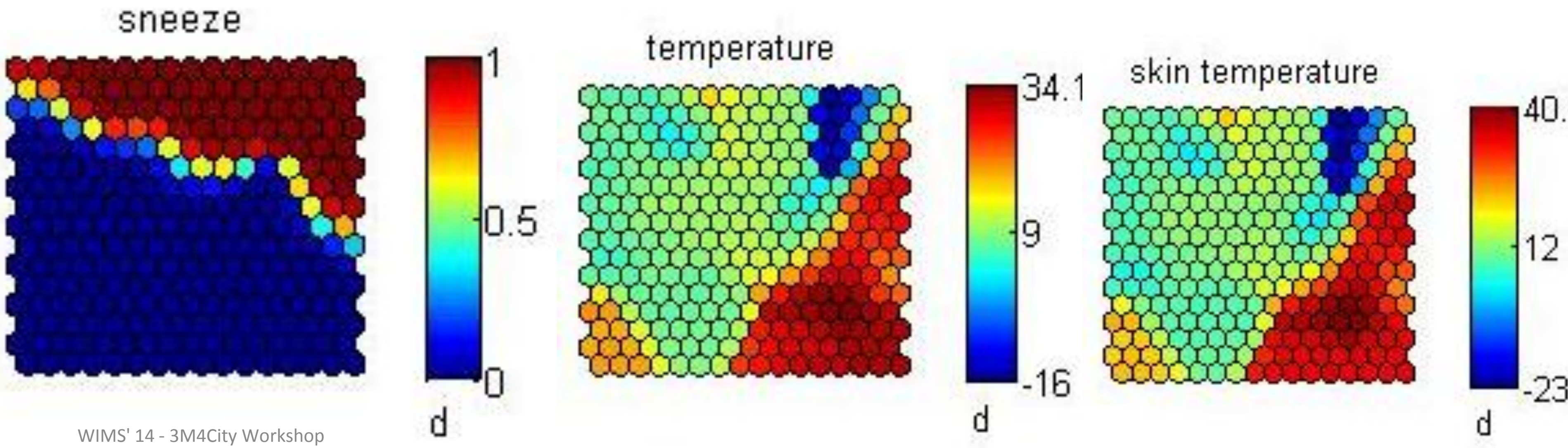




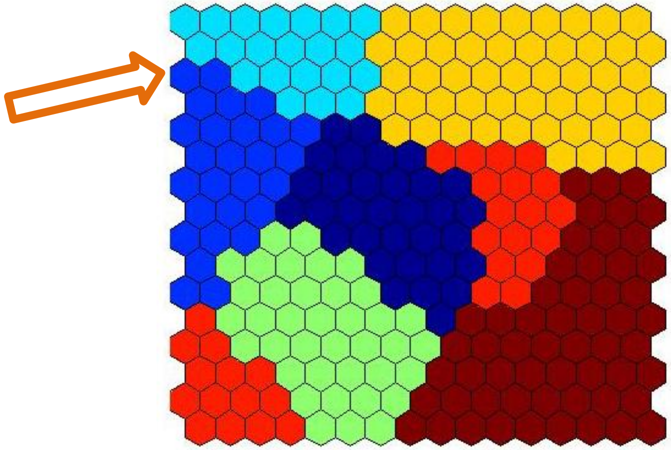


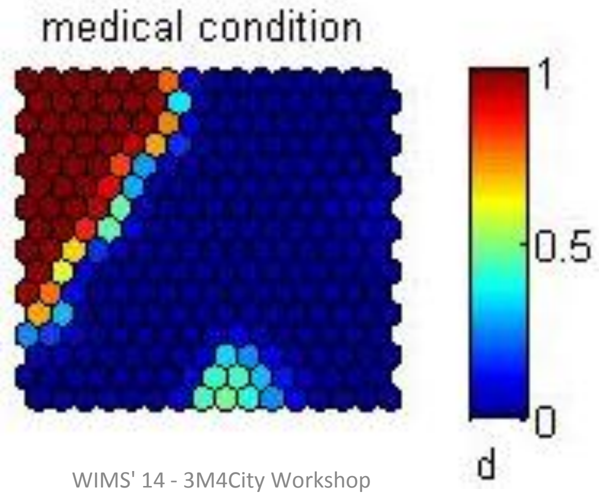
C. Relations of Sets of Words & Official Observations

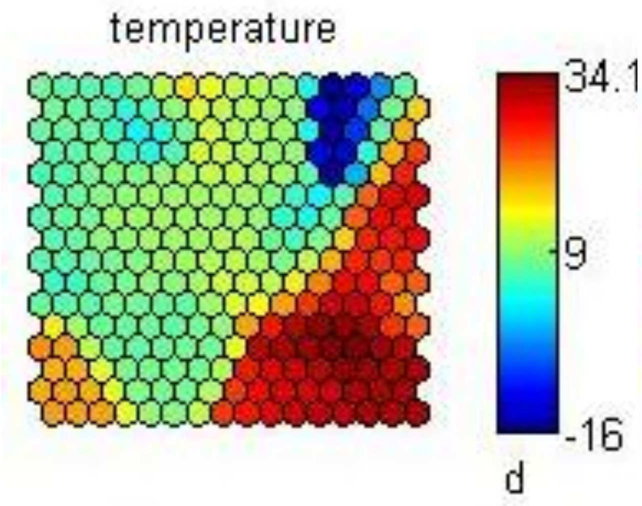


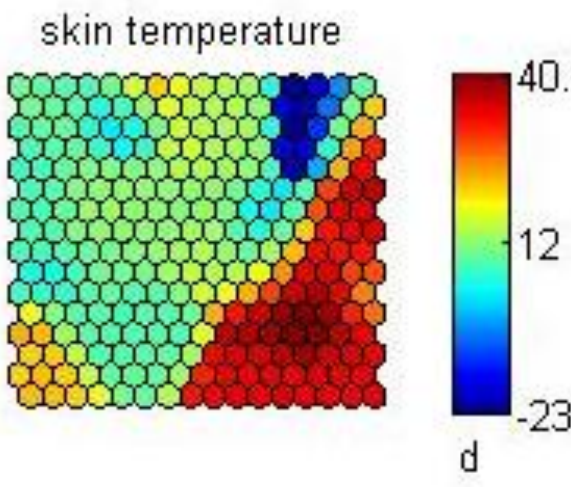


C. Relations of Sets of Words & Official Observations




medical condition


temperature


skin temperature

Conclusions

To sum up..

Combine human's observations and official measurements

Investigate the existing relations

Positive and negative relations were defined

"There is a positive relation between what people say in social media and what conditions exist in their surrounding environment"

The benefits are..

Utilize social media as a **novel** and **timely source** of information

Move towards an efficient Participatory Sensing

Future work

Automated feature extraction

Real time event detection

Requirements of PES system

Thank you!

mriga@isag.meng.auth.gr

4 June 2014

Work partially done under the Center of International Mobility (CIMO) fellowship
and with additional guidance of: prof. Mikko Kolehmainen, Maunno Rönkkö and Markus Stocker.