Images and visualisation
Statistics Netherlands (CBS)

Edwin de Jonge

December 17, 2019
Images and visualisation

Sight is the largest and fastest perceptual input channel to our brain.

This makes consuming and generating visual images interesting as:

- a data source
- medium for analysis and communication.
Image as data source

3 cases at CBS / Statistics Netherlands

Train CNN / Deplelearnig model for:

- CPI article classification of Web Shop
- Land Use Statistics
- Solar Energy production
DAMES ROKKEN

Upgrade je rokken met rechte basics, schattige pluizjes of gestroomlijnde nietjes. Of je nu zin hebt in iets briljants, onze selectie damesrokken heeft wat je nodig hebt.

Sorteren op

Filteren & sorteren

359 items

Model Product

Midirok met pailletten
€ 39.99

Jacquardgeweven rok
€ 39.99

A-linerok
€ 19.99

Kortaf maatpak

Kortaf rok

Midi rok

Maxirok

Kokerrok

Spijkerrok
CPI image classification

Statistics Netherlands uses web scraping for CPI:

- Clothing Web Shops contains > 100,000 articles
- Use Text to classify articles
- Experimented with classification image to improve classifier.

Outcome: Text classification is good (enough). Image by itself is worse, combination would give (slight) improvement
Case 2: Land Use
Case 2: Land Use Classification

- Land Use Statistics use areal photo’s to manually classify/derive land use (> 40 categories). Idea: use deep learning to speed up the process. (currently 3 years...)

- Automatic Classifier (CNN) has accuracy > 90% for large categories, but not good enough to do everything automatically

- Current research: detect land use changes, so manual task takes much less time.
Energy from Solar Panels

Use aerial photo’s to detect solar panels, as input for solar energy production estimation.

Current status:

- Basic classification working (CNN), improving labelling of dataset by creating annotation tool.
Data Science en Visualisation?
Excellent tool for both analysis and communication:

Numerical quantities focus on expected values, graphical summaries on unexpected values.

John Tukey
### Anscombe's quartet

<table>
<thead>
<tr>
<th>Dataset 1</th>
<th></th>
<th>Dataset 2</th>
<th></th>
<th>Dataset 3</th>
<th></th>
<th>Dataset 4</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
<td>$x$</td>
<td>$y$</td>
</tr>
<tr>
<td>10</td>
<td>8.04</td>
<td>10</td>
<td>9.14</td>
<td>10</td>
<td>7.46</td>
<td>8</td>
<td>6.58</td>
</tr>
<tr>
<td>8</td>
<td>6.95</td>
<td>8</td>
<td>8.14</td>
<td>8</td>
<td>6.77</td>
<td>8</td>
<td>5.76</td>
</tr>
<tr>
<td>13</td>
<td>7.58</td>
<td>13</td>
<td>8.74</td>
<td>13</td>
<td>12.74</td>
<td>8</td>
<td>7.71</td>
</tr>
<tr>
<td>9</td>
<td>8.81</td>
<td>9</td>
<td>8.77</td>
<td>9</td>
<td>7.11</td>
<td>8</td>
<td>8.84</td>
</tr>
<tr>
<td>11</td>
<td>8.33</td>
<td>11</td>
<td>9.26</td>
<td>11</td>
<td>7.81</td>
<td>8</td>
<td>8.47</td>
</tr>
<tr>
<td>14</td>
<td>9.96</td>
<td>14</td>
<td>8.1</td>
<td>14</td>
<td>8.84</td>
<td>8</td>
<td>7.04</td>
</tr>
<tr>
<td>6</td>
<td>7.24</td>
<td>6</td>
<td>6.13</td>
<td>6</td>
<td>6.08</td>
<td>8</td>
<td>5.25</td>
</tr>
<tr>
<td>4</td>
<td>4.26</td>
<td>4</td>
<td>3.1</td>
<td>4</td>
<td>5.39</td>
<td>19</td>
<td>12.5</td>
</tr>
<tr>
<td>12</td>
<td>10.84</td>
<td>12</td>
<td>9.13</td>
<td>12</td>
<td>8.15</td>
<td>8</td>
<td>5.56</td>
</tr>
<tr>
<td>13</td>
<td>7.48</td>
<td>7</td>
<td>7.26</td>
<td>7</td>
<td>6.42</td>
<td>8</td>
<td>7.91</td>
</tr>
<tr>
<td>5</td>
<td>5.68</td>
<td>5</td>
<td>4.74</td>
<td>5</td>
<td>5.73</td>
<td>8</td>
<td>6.89</td>
</tr>
</tbody>
</table>
Anscombe’s quartet

<table>
<thead>
<tr>
<th>Statistical measure</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean of x1, x2, x3, x4</td>
<td>Same: 9</td>
</tr>
<tr>
<td>Variance of x1, x2, x3, x4</td>
<td>Same: 11</td>
</tr>
<tr>
<td>Mean of y1, y2, y3, y4</td>
<td>Same: 7.50</td>
</tr>
<tr>
<td>Variance of y1, y2, y3, y4</td>
<td>Same: 4.1</td>
</tr>
<tr>
<td>Correlatie of ds1, ds2, ds3, ds4</td>
<td>Same 0.816</td>
</tr>
<tr>
<td>Linear regression ds1, ds2, ds3, ds4</td>
<td>Same: y = 3.00 + 0.500x</td>
</tr>
</tbody>
</table>

Looks same?
Let’s plot!
Uncertainty visualisation

*What is not surrounded by uncertainty cannot be the truth,*

Richard Feynman

For official statistics, at least two reasons useful:

– Communicating accuracy
– Statistical/stochastic uncertainty

Let's view two cases of stats NL (CBS)
Diabetes incidence

– Based on a (large) health survey of statistics netherlands (CBS)
Diabetes increasing
For everyone?
Small multiples: Split in groups
Reaching measurement accuracy
Traffic casualties

Verkeersdoden

Graph showing the number of traffic casualties from 1996 to 2011, with a clear decline over the years.
Traffic casualties

Exact numbers! Mortality stats (no estimation)
Traffic casualties

Let's split in smaller regions
Traffic casualties

Plot: 3d?
Traffic casualties

Over 10% year on year changes!
Case 2: Stochastic uncertainty
User Studies show:

Non-expert users can read probability intervals!
The perception of visual uncertainty representation by non-experts
Tak, Toet, van Erp, Transactions Visualisation and computer Graphics, 2014

displaying uncertainty improves data assessment
Effect of displaying uncertainty in Line and Bar charts, Van der Laan, de Jonge, Solcer, IVAPP, 2015
Weather forecast (Dutch television)
Uncertainty Viz (density)

Matthew Kay and Jessica Hullman (2019)
Uncertainty Viz (density)

US unemployment over time

Uncertainty in what unemployment was

Uncertainty in what unemployment will be

Matthew Kay and Jessica Hullman (2019)
COMUNIKOS: Eurostat project

- Goal: guidelines in COMmunicating Uncertain Knowledge in Official Statistics

Tasks:
- Describe possible sources of uncertainty
- Visualisation Guidelines
- Methods for calculating uncertainty measures
- POC on Scanner Data