Lecture 02

Data structures and types

given by

prof. Gennady Andrienko and
prof. Natalia Andrienko
Content and objectives

• Depending on the data structure, properties of the data components, and relationships between them, different analytical tasks may be meaningful and different methods and tools for the data analysis may be useful. In this lecture, you will be acquainted with the major classes of data structures.

• Three classes of data structures, the corresponding analysis tasks, and the visual and interactive tools supporting them will be introduced in more detail.

• In the practical, you will acquire your own experience in visually analysing examples of data of these three classes.
Semantic roles of data components

a reminder
Semantic roles of data components

- **Reference**: What is described?
- **Characteristic**: What is known about it?

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth date</th>
<th>School grade</th>
<th>Address</th>
<th>Distance to school, m</th>
<th>Getting to school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>17/05/2005</td>
<td>3</td>
<td>12, Pine street</td>
<td>850</td>
<td>by bus</td>
</tr>
<tr>
<td>Julia</td>
<td>23/08/2004</td>
<td>4</td>
<td>9, Oak avenue</td>
<td>400</td>
<td>on foot</td>
</tr>
<tr>
<td>Paul</td>
<td>10/12/2005</td>
<td>2</td>
<td>56, Maple road</td>
<td>1500</td>
<td>by car</td>
</tr>
<tr>
<td>Mary</td>
<td>06/10/2003</td>
<td>5</td>
<td>71, Linden lane</td>
<td>900</td>
<td>on foot</td>
</tr>
</tbody>
</table>

Referential component (referrer)

Characteristic components (attributes)
Data may have **>1 referrers**

<table>
<thead>
<tr>
<th>Referrer 1: time</th>
<th>Referrer 2: place</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>year</td>
<td>id</td>
<td>State</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1960</td>
<td>1 Alabama</td>
<td>3256740</td>
</tr>
<tr>
<td>1960</td>
<td>2 Alaska</td>
<td>226107</td>
</tr>
<tr>
<td>1960</td>
<td>4 Arizona</td>
<td>1302161</td>
</tr>
<tr>
<td>1960</td>
<td>5 Arkansas</td>
<td>1796272</td>
</tr>
<tr>
<td>1960</td>
<td>6 California</td>
<td>15717204</td>
</tr>
<tr>
<td>1960</td>
<td>8 Colorado</td>
<td>1753947</td>
</tr>
<tr>
<td>1960</td>
<td>9 Connecticut</td>
<td>2353234</td>
</tr>
<tr>
<td>1960</td>
<td>10 Delaware</td>
<td>445292</td>
</tr>
<tr>
<td>1960</td>
<td>11 District of Columbia</td>
<td>763956</td>
</tr>
<tr>
<td>1960</td>
<td>12 Florida</td>
<td>4951560</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>1972</td>
<td>54 West Virginia</td>
<td>1781000</td>
</tr>
<tr>
<td>1972</td>
<td>55 Wisconsin</td>
<td>4520000</td>
</tr>
<tr>
<td>1972</td>
<td>56 Wyoming</td>
<td>3450000</td>
</tr>
<tr>
<td>1973</td>
<td>1 Alabama</td>
<td>3393000</td>
</tr>
<tr>
<td>1973</td>
<td>2 Alaska</td>
<td>3300000</td>
</tr>
<tr>
<td>1973</td>
<td>4 Arizona</td>
<td>2058000</td>
</tr>
<tr>
<td>1973</td>
<td>5 Arkansas</td>
<td>2037000</td>
</tr>
<tr>
<td>1973</td>
<td>6 California</td>
<td>20601000</td>
</tr>
<tr>
<td>1973</td>
<td>8 Colorado</td>
<td>2437000</td>
</tr>
<tr>
<td>1973</td>
<td>9 Connecticut</td>
<td>3076000</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>2000</td>
<td>44 Rhode Island</td>
<td>1048319</td>
</tr>
<tr>
<td>2000</td>
<td>45 South Carolina</td>
<td>4012012</td>
</tr>
<tr>
<td>2000</td>
<td>46 South Dakota</td>
<td>754844</td>
</tr>
<tr>
<td>2000</td>
<td>47 Tennessee</td>
<td>5689283</td>
</tr>
<tr>
<td>2000</td>
<td>48 Texas</td>
<td>20851820</td>
</tr>
<tr>
<td>2000</td>
<td>49 Utah</td>
<td>2233169</td>
</tr>
<tr>
<td>2000</td>
<td>50 Vermont</td>
<td>608827</td>
</tr>
<tr>
<td>2000</td>
<td>51 Virginia</td>
<td>7078515</td>
</tr>
<tr>
<td>2000</td>
<td>53 Washington</td>
<td>5894121</td>
</tr>
<tr>
<td>2000</td>
<td>54 West Virginia</td>
<td>1808344</td>
</tr>
<tr>
<td>2000</td>
<td>55 Wisconsin</td>
<td>5363675</td>
</tr>
<tr>
<td>2000</td>
<td>56 Wyoming</td>
<td>493782</td>
</tr>
</tbody>
</table>
References are not always in columns

Referrer 1: place

Referrer 2: time

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>E05000026 Abbey</td>
<td>279</td>
<td>303</td>
<td>279</td>
<td>300</td>
<td>243</td>
<td>255</td>
<td>209</td>
<td>215</td>
<td>221</td>
</tr>
<tr>
<td>E05000027 Allt</td>
<td>79</td>
<td>84</td>
<td>86</td>
<td>81</td>
<td>90</td>
<td>97</td>
<td>89</td>
<td>95</td>
<td>97</td>
</tr>
<tr>
<td>E05000029 Becontree</td>
<td>96</td>
<td>89</td>
<td>105</td>
<td>106</td>
<td>112</td>
<td>110</td>
<td>101</td>
<td>102</td>
<td>103</td>
</tr>
<tr>
<td>E05000029 Chadwell Heath</td>
<td>118</td>
<td>133</td>
<td>157</td>
<td>157</td>
<td>153</td>
<td>136</td>
<td>127</td>
<td>129</td>
<td>112</td>
</tr>
<tr>
<td>E05000030 East Green</td>
<td>75</td>
<td>86</td>
<td>76</td>
<td>31</td>
<td>82</td>
<td>87</td>
<td>85</td>
<td>84</td>
<td>67</td>
</tr>
<tr>
<td>E05000031 Estoril</td>
<td>100</td>
<td>107</td>
<td>109</td>
<td>119</td>
<td>105</td>
<td>134</td>
<td>121</td>
<td>106</td>
<td>97</td>
</tr>
<tr>
<td>E05000032 Gascoigne</td>
<td>277</td>
<td>239</td>
<td>220</td>
<td>190</td>
<td>170</td>
<td>162</td>
<td>133</td>
<td>123</td>
<td>107</td>
</tr>
<tr>
<td>E05000033 Gomeldon</td>
<td>75</td>
<td>75</td>
<td>99</td>
<td>103</td>
<td>99</td>
<td>88</td>
<td>89</td>
<td>93</td>
<td>94</td>
</tr>
<tr>
<td>E05000033 Heath</td>
<td>104</td>
<td>107</td>
<td>104</td>
<td>117</td>
<td>115</td>
<td>126</td>
<td>110</td>
<td>107</td>
<td>93</td>
</tr>
<tr>
<td>E05000035 London Bridge</td>
<td>77</td>
<td>67</td>
<td>82</td>
<td>89</td>
<td>71</td>
<td>69</td>
<td>77</td>
<td>73</td>
<td>75</td>
</tr>
<tr>
<td>E05000035 Maybrook</td>
<td>95</td>
<td>74</td>
<td>72</td>
<td>92</td>
<td>95</td>
<td>102</td>
<td>96</td>
<td>100</td>
<td>100</td>
</tr>
<tr>
<td>E05000037 Parsloes</td>
<td>77</td>
<td>79</td>
<td>70</td>
<td>92</td>
<td>95</td>
<td>97</td>
<td>92</td>
<td>89</td>
<td>94</td>
</tr>
<tr>
<td>E05000033 River</td>
<td>112</td>
<td>103</td>
<td>114</td>
<td>127</td>
<td>115</td>
<td>109</td>
<td>99</td>
<td>95</td>
<td>86</td>
</tr>
<tr>
<td>E05000033 Thames</td>
<td>263</td>
<td>238</td>
<td>254</td>
<td>237</td>
<td>215</td>
<td>208</td>
<td>180</td>
<td>178</td>
<td>175</td>
</tr>
<tr>
<td>E05000040 Vicenza</td>
<td>81</td>
<td>75</td>
<td>83</td>
<td>87</td>
<td>88</td>
<td>97</td>
<td>99</td>
<td>76</td>
<td>80</td>
</tr>
<tr>
<td>E05000041 Village</td>
<td>111</td>
<td>124</td>
<td>143</td>
<td>129</td>
<td>115</td>
<td>134</td>
<td>130</td>
<td>129</td>
<td>133</td>
</tr>
<tr>
<td>E05000042 Whitchurch</td>
<td>109</td>
<td>93</td>
<td>96</td>
<td>97</td>
<td>114</td>
<td>115</td>
<td>107</td>
<td>107</td>
<td>93</td>
</tr>
<tr>
<td>E05000043 Brunswick Park</td>
<td>70</td>
<td>88</td>
<td>94</td>
<td>79</td>
<td>88</td>
<td>81</td>
<td>56</td>
<td>54</td>
<td>58</td>
</tr>
<tr>
<td>E05000044 Burnt Oak</td>
<td>88</td>
<td>108</td>
<td>102</td>
<td>114</td>
<td>95</td>
<td>93</td>
<td>79</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>E05000045 Childs Hill</td>
<td>119</td>
<td>125</td>
<td>129</td>
<td>142</td>
<td>136</td>
<td>127</td>
<td>118</td>
<td>106</td>
<td>105</td>
</tr>
<tr>
<td>E05000046 Colindale</td>
<td>93</td>
<td>114</td>
<td>114</td>
<td>107</td>
<td>95</td>
<td>93</td>
<td>76</td>
<td>76</td>
<td>76</td>
</tr>
<tr>
<td>E05000047 Coppetts</td>
<td>99</td>
<td>104</td>
<td>115</td>
<td>113</td>
<td>100</td>
<td>103</td>
<td>94</td>
<td>97</td>
<td>91</td>
</tr>
<tr>
<td>E05000048 East Barnet</td>
<td>81</td>
<td>62</td>
<td>80</td>
<td>68</td>
<td>82</td>
<td>81</td>
<td>66</td>
<td>65</td>
<td>61</td>
</tr>
<tr>
<td>E05000049 East Finchley</td>
<td>85</td>
<td>87</td>
<td>95</td>
<td>92</td>
<td>81</td>
<td>84</td>
<td>59</td>
<td>67</td>
<td>61</td>
</tr>
<tr>
<td>E05000050 Edgware</td>
<td>107</td>
<td>108</td>
<td>107</td>
<td>121</td>
<td>114</td>
<td>106</td>
<td>89</td>
<td>84</td>
<td>96</td>
</tr>
<tr>
<td>E05000051 Finchley</td>
<td>63</td>
<td>55</td>
<td>58</td>
<td>64</td>
<td>52</td>
<td>56</td>
<td>51</td>
<td>53</td>
<td>56</td>
</tr>
<tr>
<td>E05000052 Garden Suburb</td>
<td>87</td>
<td>83</td>
<td>88</td>
<td>105</td>
<td>70</td>
<td>72</td>
<td>74</td>
<td>65</td>
<td>66</td>
</tr>
<tr>
<td>E05000053 Golders Green</td>
<td>94</td>
<td>98</td>
<td>101</td>
<td>112</td>
<td>95</td>
<td>96</td>
<td>79</td>
<td>70</td>
<td>83</td>
</tr>
<tr>
<td>E05000054 Hale</td>
<td>56</td>
<td>63</td>
<td>65</td>
<td>70</td>
<td>72</td>
<td>70</td>
<td>54</td>
<td>54</td>
<td>55</td>
</tr>
<tr>
<td>E05000055 Hendon</td>
<td>98</td>
<td>83</td>
<td>101</td>
<td>94</td>
<td>106</td>
<td>86</td>
<td>84</td>
<td>93</td>
<td>92</td>
</tr>
<tr>
<td>E05000056 High Barnet</td>
<td>79</td>
<td>92</td>
<td>105</td>
<td>109</td>
<td>101</td>
<td>75</td>
<td>80</td>
<td>80</td>
<td>72</td>
</tr>
<tr>
<td>E05000057 Mill Hill</td>
<td>79</td>
<td>90</td>
<td>103</td>
<td>93</td>
<td>85</td>
<td>90</td>
<td>84</td>
<td>74</td>
<td>71</td>
</tr>
<tr>
<td>E05000059 Oakleigh</td>
<td>57</td>
<td>63</td>
<td>72</td>
<td>78</td>
<td>72</td>
<td>61</td>
<td>58</td>
<td>62</td>
<td>51</td>
</tr>
</tbody>
</table>
**Data components viewed as variables**

- **Referrers** ~ independent variables
- **Attributes** ~ dependent variables
- Data represent (are generated by) a *function*
  \[ \text{Referrers} \rightarrow \text{Attributes} \]
  - References (values of referrers): function’s inputs
  - Characteristics (values of attributes): function’s outputs
- Function’s *behaviour*:
  how the outputs vary over the set of inputs \( \rightarrow \)
  how the characteristics vary over the set of references
  - E.g., how the crime rates vary over space and time
Definitions of data types and structures
Types and properties of data components

Types of values:
- Numeric
- Textual
  - Predefined values (e.g., codes)
  - Free text
- Spatial
  - Coordinates
  - Place names
  - Addresses
- Temporal
- Other (image, video, audio, ...)

Scales of measurement*:
- Nominal (¬ order, ¬ distances)
  - gender, nationality, ...
- Ordinal (√ order, ¬ distances)
  - evaluations: bad, fair, good, excellent
- Interval (√ order, √ distances, ¬ ratios,
  ¬ meaningful zero)
  - temperature, time, ...
- Ratio (√ order, √ distances, √ ratios,
  √ meaningful zero)
  - quantities, distances, durations, ...

<table>
<thead>
<tr>
<th>Name</th>
<th>Birth date</th>
<th>School grade</th>
<th>Address</th>
<th>Distance to school, m</th>
<th>Getting to school</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peter</td>
<td>17/05/2005</td>
<td>3</td>
<td>12, Pine street</td>
<td>850</td>
<td>by bus</td>
</tr>
<tr>
<td>Julia</td>
<td>23/08/2004</td>
<td>4</td>
<td>9, Oak avenue</td>
<td>400</td>
<td>on foot</td>
</tr>
<tr>
<td>Paul</td>
<td>10/12/2005</td>
<td>2</td>
<td>56, Maple road</td>
<td>1500</td>
<td>by car</td>
</tr>
<tr>
<td>Mary</td>
<td>06/10/2003</td>
<td>5</td>
<td>71, Linden lane</td>
<td>900</td>
<td>on foot</td>
</tr>
</tbody>
</table>

Essential properties of sets (value domains of data components)*

- **Ordering** between the elements
  - No ordering, partial ordering, full (linear) ordering

- Existence of *distances* between the elements
  - No distances, has distances

- **Continuity**
  - Discrete, continuous

- For any data component, there is a certain set (domain) of possible values

  ⇒ Data components can be characterised in terms of the properties of their domains

---

Three main types of referrers according to their value domains*

- **Time**
  - Has linear ordering, has distances, continuous

- **Space (2D, 3D)**
  - No ordering, has distances, continuous

- **Population** *(in a general sense, like statistical population, i.e., any set of material or abstract objects)*
  - No ordering, no distances, discrete
    - Further will be referred to as “discrete objects”, or simply “objects”, to avoid confusion with “population” in demographic sense

---

Main classes of data structures

according to the types of the referrers

• **Object-referenced data:**
  • attribute values refer to discrete objects

• **Time-referenced data**, a.k.a. **time series:**
  • attribute values refer to different times (moments or intervals)

• **Space-referenced data**, a.k.a. **spatial data:**
  • attribute values refer to different spatial locations (points, lines, areas, volumes in 3D space)

• **Object-referenced time series:**
  • attribute values refer to discrete objects and to different times (i.e., the data have 2 referrers)

• **Spatial time series:**
  • attribute values refer to different spatial locations and to different times (i.e., the data have 2 referrers)
Multi-dimensional data (a.k.a. multivariate data)

- Data including multiple attributes
- May have various types of references: discrete objects, spatial locations, times, or combinations of these
  - Multiple attributes referring to times: multidimensional (multivariate) time series
  - Multiple attributes referring to places: multidimensional spatial data
  - Multiple attributes referring to places + times: multidimensional spatial time series
An example: London crime data

• Referrers:
  • London wards – spatial
  • Financial years – temporal

• Attributes:
  • 10 crime rates (total + different types of crimes)
  • 10 crime counts (total + different types of crimes)

⇒ Multidimensional spatial time series
Types of objects

*based on their properties* and attributes**

• **Generic objects**
  • Wine varieties, car models, plant specimens, ...

• **Spatial objects**
  • Have locations in space ⇒ the attributes include the location
    • Districts, buildings, streets, rivers, ...

• **Temporal objects**, a.k.a. *events*
  • Have limited existence time ⇒ the attributes include the time of existence
    • Instant objects: have no duration; only the appearance time needs to be specified
      • Tweet postings, bank transactions, lightning strokes ...
    • Durable objects: have duration; the attributes need to include the time of appearance + the time of disappearance or the duration
      • Holidays, electoral campaigns, classes, breaks, TV shows, ...

* What is observed in the reality  ** Representations of properties in the data
Types of objects (continued)

spatio-temporal objects

• Spatial events
  • Events that have location in space \( \Rightarrow \) the attributes include the spatial location and the existence time (instance or interval)
  • Lightnings, geolocated tweet postings, earthquakes, traffic jams, ...

• Moving objects
  • Spatial objects whose locations change over time \( \Rightarrow \) the attributes include spatial locations at different times, i.e., time series of spatial locations, called \textit{trajectories}
  • People, animals, vehicles, storms, oil spills, ...

• Trajectories can be treated as spatio-temporal objects
  • Spatial location = spatial footprint
  • Existence time = time from the beginning till the end of the movement
  • Other properties \( \rightarrow \) attributes: shape, travelled distance, mean speed
Relationships between types of objects

Venn diagram
Classes of data to be considered in this lecture

• Object-referenced attributes
  • References: generic objects (have no temporal or spatial properties, or these properties are ignored)

• Space-referenced attributes
  • References: spatial locations or stationary spatial objects

• Spatial event data
  • References: spatial events

➢ All these are data with single referrers

➢ Next lecture: data with 2 referrers
Questions?

Definitions of data types and structures
Following material

• Classes of data
  • Object-referenced attributes
  • Space-referenced attributes
  • Spatial event data

• For each class of data, we shall consider the typical synoptic tasks and the methods to accomplish them

• Synoptic tasks (a reminder):
  • Distinguishing property: a set or subset of data items is considered in its entirety; individual items are not of interest.
  ⇒ Synoptic tasks entail abstraction from individual items to sets
  ⇒ Synoptic tasks require analysis methods and visualisations supporting abstraction
Major classes of synoptic tasks: a reminder

• Describe the behaviour of one or more attributes

• Find subsets of references where attributes have particular behaviours

• Compare two or more behaviours (find similarities and differences)
  • Different attributes over the same set of references
  • Same attributes over different subsets of references

• Relate behaviours of two or more attributes
Classes of synoptic tasks with examples

• Describe the behaviour of one or more attributes
  • Describe the variation of the crime rates over time and space

• Find subsets of references where attributes have particular behaviours
  • Find time periods of decreasing burglary rates

• Compare two or more behaviours (find similarities and differences)
  • Different attributes over the same set of references
    • Compare the variations of the burglary and robbery rates
  • Same attributes over different subsets of references
    • Compare the temporal variations of the burglary rates in the eastern and western parts

• Relate behaviours of two or more attributes
  • Relate the crime rates to socio-demographic characteristics
Analysis of object-referenced data
Attributes referring to discrete objects
Major classes of synoptic tasks

• Describe the behaviour of one or more attributes →
  Describe the distribution of the attribute values and value combinations over the set of objects

• Find subsets of references where attributes have particular behaviours →
  Find subsets (groups) of objects with particular attribute values or combinations, or with value combinations substantially differing from the bulk

• Compare two or more behaviours →
  Compare two or more value distributions
  • Different attributes over the same set of objects
  • Same attributes over different subsets of objects

• Relate behaviours of two or more attributes →
  Relate value distributions of two or more attributes
Task: Describe the distribution of a single numeric attribute

A possible approach: bar diagram (sorted). However, for supporting synoptic tasks, bars for all objects must be seen simultaneously \(\Rightarrow\) it is limited to quite a small number of objects.
Task: Describe the distribution of a single numeric attribute

A better option: frequency histogram

Shows the most popular value intervals and exhibits outliers. The shape can be compared with the “model” distribution shapes defined in statistics.
A frequency histogram example

The most frequent values

Frequent values

Outliers

Becontree; Eastbrook; Eastbury; Longbridge; Mayesbrook; Thames; Village; Whalebone; Brunswick Park; Childs Hill ... (198 in total)
Frequency histogram for a qualitative (nominal) attribute

May be better when sorted
Other synoptic tasks supported by interactive frequency histograms

Compare value distributions of several attributes over the same object set.
Other synoptic tasks supported by interactive frequency histograms

Relate value distributions of several attributes.
Task: Relate value distributions of two attributes

Possible relationships: positive correlation, negative correlation, independence; clusters (groups of similar value combinations).
Task: Describe joint value distribution of multiple attributes

Step 1: projection (dimensionality reduction)

Combinations of values of multiple attributes

Positions in 2D space

Projection method here: Sammon’s mapping; may be also MDS, PCA, …
Task: Describe joint value distribution of multiple attributes

Step 2: superimpose the projection on a continuous 2D colour scale

A parallel coordinates plot helps to understand what value combinations correspond to different areas of the projection plot.
Tasks: describe joint value distribution of multiple attributes; find groups of objects with similar value combinations

Select (by clicking or dragging) a group of close points in the projection plot; then look at the corresponding marks in the PCP and in other displays.
Further use of projection

Grouping of objects by projection space tessellation

Creates object classes, which can be propagated to other displays.
Task: find groups of objects with particular value combinations

Supporting interactive tool: filtering by attribute values.
Techniques for analysing object-referenced data

- **Visualisations**: bar diagram, parallel coordinates plot (both suitable for small number of objects), frequency histogram, scatter plot, 2D histogram, projection plot.

- **Transformations**: aggregation (used in histograms), projection (dimensionality reduction).
  - Aggregation: helps to deal with numerous objects; supports abstraction.
  - Projection: helps to deal with multiple attributes.

- **Interactive operations**: selection, grouping (classification), filtering.
  - Support relating multiple attributes and finding groups of objects with particular/ close/ outstanding value combinations.
Questions?

Analysis of object-referenced data
Analysis of space-referenced data

Attributes referring to spatial locations or spatial objects
Space as a special reference domain

• Essential property: distances between elements
  ⇒ Spatial locations and spatial objects have neighbourhoods

• Behaviour of attributes over space = spatial distribution of attribute values and combinations

• In analysing and describing spatial distributions, spatial neighbourhood is taken into account

• Types (patterns) of spatial distribution based on neighbourhood
  • Spatial clusters: groups of neighbouring locations or objects having similar attribute values
  • Spatial smoothness: attribute values of neighbouring locations or objects do not differ much
Tobler's first law of geography

• Everything is related to everything else, but near things are more related than distant things*.

• A.k.a. *spatial dependence*.

• Implication: neighbouring objects or locations are expected to have similar attribute values.
  ⇒ Spatially smooth or spatially clustered attribute values are a kind of “normal” spatial distribution
    • *Spatial outliers*: locations or objects whose attribute values differ much from those of the neighbours.

• However, the first law of geography may not keep for spatially aggregated data (e.g., by districts).

Spatial directions

• Space has no natural ordering between locations.

• However, introducing a coordinate system also defines directions, such as compass directions in the geographic space.

• Type (pattern) of spatial distribution based on directions:
  • Spatial trend: increase or decrease of attribute values
    • in some direction w.r.t. the coordinate system (e.g., from north to south) or
    • in all/many directions from some location (e.g., from the city centre to the periphery)
Geographic space

• Spatially referenced data typically refer to geographic rather than abstract space.

• The geographic space is \textit{heterogeneous}: consists of land and sea, plains, mountains, and valleys, rivers, lakes, and deserts, forests and built areas, ...

• The spatial distribution of attribute values may be affected by the variation of the properties of the underlying geographic space.
  • This interferes with the first law of geography: near things may be less related or less similar due to the differences of the underlying space.
Distribution of attribute values over geographic space

• In analysing spatial distributions, the geographic background should be taken into account.

• Types (patterns) of spatial distribution related to heterogeneous geographic background:
  • Differentiation based on altitude, land cover, land use, etc.
  • Alignment of similar values (elongated spatial cluster) or trend (increase or decrease) along a linear natural or man-made feature (river, road, ...)
  • Spatial trend towards/outwards a boundary line (e.g., a coastline)
  • Concentration of high/low/specific values around some feature (e.g., a factory polluting air)
Major classes of synoptic tasks

• Describe the behaviour of one or more attributes →
  Describe the *spatial distribution* of the attribute values and value combinations

• Find subsets of references where attributes have particular behaviours →
  Find parts of space with particular attribute values or combinations, or with value combinations substantially differing from the bulk

• Compare two or more behaviours →
  Compare two or more spatial distributions of attribute values
    • Different attributes over the same part of space (territory)
    • Same attributes over different parts of space

• Relate behaviours of two or more attributes →
  Relate value distributions of two or more attributes
Map: the main instrument in analysing spatial data

- A map can show spatial distribution of spatial objects and space-referenced attribute values.
  - It allows the analyst to observe the spatial distribution patterns based on distance/neighborhood relationships and on spatial directions.

- For geographic space, a map can also show the heterogeneous geographic background.
  - It thereby allows the analyst to observe the spatial distribution patterns based on relationships with the background properties and features.

Cartograms:
- distort the spatial directions
- distort the distance and (often) neighborhood relationships
- the geographic background can hardly be shown together with data
Efficient maps of spatial distributions of attribute values

The most efficient map of a spatial distribution is a map from which the distribution is perceived as a single image, i.e., “in the minimum instant of vision” (J.Bertin).

The use of colour darkness (above) or colour hue (left) creates an efficient map of the spatial distribution of a single numeric (above) or qualitative (left) attribute.
Each diagram on the map requires an individual instance of perception, i.e., $N$ of images = $N$ of diagrams = $N$ of districts.

Recall from lecture 1:

Each map requires one instance of perception, i.e., $N$ of images = 2.
Limitation regarding the number of attributes

• Bertin: an image can be composed using at most two display dimensions and one retinal visual variable.

• In a map, the two display dimensions are used for representing spatial positions.

  ⇒ Attribute values can only be represented by retinal variables: size, value (colour darkness), colour (hue), orientation, shape, or texture.

  ⇒ Creating a single image is only possible for a single space-referenced attribute ...

    • ... or several attributes need to be somehow integrated into a single attribute.
Choropleth map (a reminder)

Values of a numeric attribute are represented by the visual variable ‘value’ (darkness). Darker shades correspond to higher attribute values. The shades are used for painting areas or objects on the map. The spatial distribution is perceived as a single image.
A diverging scale of colour shades uses two distinct colour hues for representing positive and negative values or positive and negative differences from a chosen central value, such as the overall mean. Darker shades correspond to larger differences. The map is perceived as a single image and exposes spatial clusters of high and low values.
Map with proportional symbols (bars)
Map with proportional symbols (bars)

Diverging bars use orientation (upward and downward) and, complementarily, two distinct colour hues for showing positive and negative attribute values or differences from a selected central value, such as the overall mean.
Map with proportional symbols (circles)

Problem: too much visual clutter; requires zooming for discerning the symbols.
Map with proportional circles

The symbols are now discernible, but the overall view is lost.
Map with proportional circles, diverging scale

Similarly to bars, circles of two distinct colour hues can show positive and negative values or differences (the orientation cannot be used in this case).
Choropleth maps vs. proportional symbol maps

- Choropleth maps and maps with proportional symbols are used for representing the spatial distribution of a single numeric attribute.

- Advantages of a choropleth map:
  - provides a single image and thereby supports synoptic tasks;
  - is free from overlapping and clutter;
  - small differences in shades may be easier detectable than small differences in sizes.

- Maps with proportional symbols can also be efficient when they are not visually cluttered.

- Both choropleth maps and proportional symbol maps become more expressive and effective when diverging scales are applied.
Mapping of qualitative attribute values

Qualitative attribute values can best be represented by the retinal visual variable ‘colour’ (hue).
Synoptic tasks supported by maps alone

- Describe the spatial distribution of values of a single attribute.

- Compare several spatial distributions:
  - Different attributes over the same part of space (territory)
  - Same single attribute over different parts of space
  - Comparisons are supported by several juxtaposed maps (“small multiples” technique).

- For other tasks, maps need to be combined with other display types and interaction techniques.
Combining maps with other techniques

• All visualisation and interaction techniques supporting analysis of object-referenced data can be applied to space-referenced data.

• The specific tasks of spatial data analysis require these techniques to be used in combination with maps.

• Combining information from different views can be supported by interactive techniques for display linking:
  • Selection
  • Classification
Two possible approaches to linking maps with attribute displays

• Attribute displays → map
  • Select or classify objects on a plot, diagram, or histogram and observe the spatial distribution of the selected objects or object classes on a map
    • Are there any spatial patterns?

• Map → attribute displays
  • Select or classify spatial objects on a map based on their spatial positions and observe on the attribute displays what values and value combinations occur in the selected part of space or are characteristic for the different parts of space (i.e., spatial classes of objects).
    • Are there any distinguishing value subsets or combinations?
Selection: attribute displays → map

Attribute displays (frequency histograms): we have selected districts with relatively high percentages of people having bad or very bad health.

Map: shows us that the districts where many people have bad or very bad health are grouped in spatial clusters.
Selection: map → attribute displays

Map: we have selected the districts on the west of London.

Attribute displays show us that the western districts tend to relatively high proportions of people having good health, average proportions of those having fair health, and lower proportions of people with bad or very bad health.
We see that the spatial distribution of the attribute “average distance travelled to work” has a spatial trend of increasing values from the centre to the periphery. This could also be seen on an unclassified choropleth map; however, classes can be propagated to other displays.
% of female population by economic activity
Classification based on multiple attributes (through projection) → Map

The map exhibits a spatial trend (centre to periphery) in the spatial distribution of the age structure represented by 16 numeric attributes.
Interpreting the colours

PCP; the classes are summarised by quantiles

The classes to view can be interactively selected.
Space-based interactive classification on a map → attribute displays
Techniques for analysing space-referenced data

• Describe the spatial distribution of a single attribute; compare spatial distributions of singular attributes
  ➢ Maps with colouring
    • Retinal variable ‘value’ (darkness): numeric or ordinal attributes
    • Retinal variable ‘colour’ (hue): qualitative attributes
    • Several juxtaposed maps ("small multiples" technique) support comparisons of spatial distributions

• Describe the spatial distribution of value combinations of multiple attributes; find parts of space with particular or outstanding value combinations; relate attribute behaviours
  ➢ Maps + attribute displays linked by interactive techniques
    • Selection
    • Classification
Questions?

Analysis of space-referenced data
Analysis of spatial events
Objects having spatial locations and existence times
Spatial event data

• Spatial event data structure:
  • 1 referrer: set of objects
  • 2 mandatory attributes: *spatial location* + *time of existence*
  • any other attributes, further called *thematic attributes*

• Spatial location and existence time are *attributes* of the objects
  ⇒ The main synoptic tasks address the behaviour (distribution) of the spatial locations and existence times over the set of objects.

• However, space and time can also be considered as independently existing *containers* of the objects.
  ⇒ The main synoptic tasks may be re-formulated to address the distribution of the objects over the space and time, i.e., the *spatio-temporal distribution* of objects
Space and time as object containers

• Considering space and time as containers of objects is quite intuitive.
  • We can easily imagine space and time without objects. It is difficult to do
    the same for other attributes (e.g., size).

• In visualisation, it is typical to represent space and time by display
  dimensions and objects by marks located within the display space.
  • I.e., the display conveys the idea of the objects being contained in space
    and/or time.
  • This representation is usual for people and therefore easily
    understandable.

➢ We take this *absolute view* of time and space as object containers
  and formulate the synoptic tasks accordingly.
Major classes of synoptic tasks

1) Attributes: space and time

• Describe the behaviour of one or more attributes →
  Describe the spatio-temporal distribution of the events

• Find subsets of references where attributes have particular behaviours →
  Find subset of objects forming particular spatio-temporal patterns, such as spatio-temporal concentration (cluster), alignment, ...

• Compare two or more behaviours →
  Compare spatio-temporal distributions of event (sub)sets
  • Different events (by thematic attributes) in the same space and time
  • Events from different time periods in the same space
  • Events in different parts of space

• Relate behaviours of two or more attributes →
  Relate the spatial distribution to time and vice versa
Major classes of synoptic tasks

2) Attributes: thematic attributes

• Spatial events are *objects*.
  ⇒ Spatial event data are a special case of *object-referenced data*.
  ⇒ All synoptic tasks defined for object-referenced data apply, in particular, to spatial event data.
    • I.e., tasks addressing the distribution of the attribute values over the set of objects.
  ⇒ All techniques supporting the analysis of object-referenced data can be used for spatial event data.

• Spatial events are *spatial objects*.
  ⇒ Spatial event data are a special case of *space-referenced data*.
  ⇒ All synoptic tasks defined for space-referenced data apply, in particular, to spatial event data.
    • I.e., tasks addressing the spatial distribution of the attribute values.
  ⇒ All techniques supporting the analysis of space-referenced data can be used for spatial event data.
Major classes of synoptic tasks

2) Attributes: thematic attributes (continued)

• Additionally, spatial events are *spatio-temporal objects*.
  ⇒Values of the thematic attributes refer to spatial locations and time moments and intervals.
  ⇒It is appropriate to consider the behaviour of the attribute values over space and time, i.e., the *spatio-temporal distribution of the attribute values*.

• The tasks are formulated similarly to the tasks addressing the spatio-temporal distribution of the events.
Visual displays of spatial events

Map: shows the spatial distribution

Space-time cube*: represents 2D space + time as a single 3D continuum, in which the objects (spatial events) are positioned.

Temporal display (e.g., frequency histogram of the occurrence times): shows the temporal distribution

Spatial distribution of events

Point events (i.e., having no spatial extents) are typically represented on a map by dot symbols (small circles). Very often the dots overlap ⇒ semi-transparent drawing is recommended. The degree of transparency is adjusted to the dot density. Such a way of rendering helps us to observe the variation of the event density over the territory and detect spatial clusters of events.
Investigation of the spatial variation of the event density can be supported by spatial aggregation of the events, e.g., by cells of a regular grid (here: 1×1 km). For each cell, the number of events is counted; additionally, thematic attributes can be summarized by computing the mean, mode, median, etc. Resulting data type: space-referenced attributes (considered before).
Spatial aggregation by arbitrary areas

Events can also be aggregated by arbitrary areas, such as administrative districts (e.g., to relate event numbers and/or attributes to other characteristics of the areas). However, arbitrary areas may significantly differ in sizes, which distorts the perception of the event density. ⇒ Choropleth maps, like this one, should not be used for representing counts by areas.
Aggregation: absolute counts transformed to relative

A correct image of the variation of the event density can be obtained by transforming the absolute event counts to relative w.r.t. the areas of the districts. When the events are related to lives and activities of people (e.g., tweets), it can also make sense to compute the relative counts w.r.t. the population of the districts.
A frequency histogram of the event occurrence times shows the overall temporal distribution of the events and exhibits temporal patterns such as temporal trends and periodicity. A periodic temporal distribution can be additionally explored using a 2D histogram, e.g., with the dimensions corresponding to the days of the week and hours of the day.
A static perspective view of a 3D representation is insufficient for observing and exploring the content. Interactive operations for changing the perspective (rotation, shifting, and tilting) are necessary.
A space-time cube (STC) can exhibit several patterns of spatio-temporal distribution (types of spatio-temporal behaviour) appearing as horizontal “layers” and/or “gaps” (= periods of high and low event density), vertical “columns” (= high event density in some area for a long time), and “lumps” (= spatio-temporal clusters, i.e., groups of events that occurred closely in space and time).
Spatio-temporal distribution of events

Diagonally elongated clusters may mean that the events were caused by a moving phenomenon, such as a storm or parade. This example: lightning events.
Spatial, temporal, and spatio-temporal distribution of thematic attribute values
Spatio-temporal distributions of different attribute values

Different attribute values are selected using interactive attribute-based filtering.
Visualisation of thematic attributes of events

Although retinal visual variables can be used for representing the values of thematic attributes of spatial events on a map and in a space-time cube, the visual clutter and overlapping of marks make the displays illegible and practically useless for the analysis.
Spatial filtering of events

Note the focuser position
Temporal filtering of events
Insufficiency of visualisation and interaction techniques

• Map and STC do not work well when the events are numerous and dense.
  • The spatial and spatio-temporal distribution of the events themselves and their thematic attributes cannot be effectively analysed due to display clutter and greatly overlapping marks.

• Besides, STC may not work well when the time span of the data is long.
  • Spatio-temporal distribution patterns are not well visible.

• Interactive filtering can only partly reduce the display clutter while eliminating the overall view.

• Data transformations and computational techniques are strongly needed.
  • E.g., computational detection of clusters – to be considered later
Spatio-temporal aggregation of events

• Spatial events and their thematic attributes can be aggregated spatially by areas (as considered before) and, simultaneously, by time intervals.

• Resulting data type: space-referenced time series of
  • event counts, densities, counts per capita, ...
  • statistical summaries of thematic attributes: mean, median, mode, minimum, maximum, quantiles, ...

• Structure of the resulting data: $S \times T \rightarrow A$
  • $S$: the areas by which the events have been aggregated
  • $T$: the time intervals by which the events have been aggregated
  • $A$: the aggregate attribute values
Spatio-temporal aggregation of events: an example

Geolocated tweets have been aggregated by cells of a regular grid and hourly time intervals. For each grid cell, there is a time series of hourly tweet counts.
Techniques for analysis of spatio-temporal event data: a summary

• Visual displays: map, space-time cube, time-based frequency histograms, various display types for thematic attributes.

• Interactive techniques: selection, classification, filtering.

• Data transformations: spatial aggregation, spatio-temporal aggregation.
Questions?

Analysis of spatial events
Summary: data structures and types

• Data components are distinguished according to their semantic roles: *referrers* and *attributes*

• Data types are distinguished according to the types and structure of the references
  • Major types: *discrete objects*, *space*, *time*
  • Structure: 1D (single referrer), 2D (2 referrers), …
    • In case of 2+D, each reference is a combination of values of the referrers; the set of possible references is a Cartesian product $R_1 \times R_2 (\times \ldots)$.

• Discrete objects are distinguished according to their attributes:
  • *Generic objects*: have arbitrary attributes
  • *Spatial objects*: have attribute ‘spatial location’
  • *Temporal objects* (events): have attribute ‘existence time’
  • *Spatio-temporal objects*: have attributes ‘spatial location’ and ‘existence time’
Summary: synoptic analysis tasks

- Data ~ a function (in mathematical sense), i.e., a mapping from the domain of possible references to the domain of possible characteristics (combinations of attribute values) $f: R \rightarrow A$
  - $R$ (referrers): independent variables
  - $A$ (attributes): dependent variables
- Synoptic analysis tasks address the *behaviour* of this function, i.e., how $A$ varies over the value domain of $R$
  - I.e., how the characteristics (attribute values or combinations) are distributed over the set of references.
- Tasks: describe the behaviour, detect particular type of behaviour (pattern), compare behaviours, relate behaviours
Summary: analytical techniques

• References: generic objects
  • Visualisations: attribute displays (frequency histogram, scatter plot, parallel coordinates)
  • Transformations: aggregation, projection (dimensionality reduction)
  • Interactions: selection, classification, attribute-based filtering

• References: space (spatial locations) or spatial objects
  • Visualisations: maps + attribute displays
    • Linking through interactive selection and classification; 2 ways of propagating selections and classes: attribute displays → map and map → attribute displays
  + All other techniques suitable for generic objects
Summary: analytical techniques (continued)

- References: spatial events (= objects having spatial locations and existence times)
  - Space (S) and time (T) can be viewed as object containers
    - Separately or in combination $S \times T = \text{space-time continuum}$
  $\Rightarrow$ Specific definition of behaviour: $\Rightarrow$ distribution of the objects and their characteristics over $S$, $T$, and $S \times T$

- Specific visualisations:
  - Map: shows the distribution over $S$
  - Temporal displays (time-based histograms, Gantt chart): show the distribution over $T$
  - Space-time cube: shows the distribution over $S \times T$

- Specific interactive techniques:
  - Spatial filtering
  - Temporal filtering

+ All techniques suitable for generic objects or spatial objects/locations
Reading:

http://dx.doi.org.wam.city.ac.uk/10.1007/3-540-31190-4

Natalia and Gennady Andrienko

Exploratory Analysis of Spatial and Temporal Data
A Systematic Approach


Chapter 2: Data
Chapter 3: Tasks