1

- **Cluster Analysis**
  - Technique of exploratory data analysis
  - Do data divide into groups or ‘clusters’
  - Are there ‘natural’ divisions into groups & sub-groups
  - Some attempt to answer questions of whether observed clusters are ‘real’ or by chance ---
    - But that needs a statistical basis
  - Most techniques based on intuitive ideas

2

- Steps needed
  - Collect data
  - Decide on measure of similarity
    - Packages offer a default & alternative choices
    - Depend on type of data:-
      continuous, binary, categorical (unordered) categorical (ordered), mixed
  - Decide on rule to ‘form’ a cluster
    - Different rules form different types
    - Loose/compact/robust (not influenced by outliers)
  - Decide on ‘suitable’ number of clusters
    - Display dendrogram (tree diagram) & visually
  - (validation of solution)

3

- **Similarity measures**
  - **Euclidean**
    - Ordinary measurements
  - Binary
    - Presence/absence
  - Manhatan
    - binary
  - Other specialist ones

4

- Clustering method
  - Single link
    - Item joins if close to one member of cluster
  - Complete
    - Joins if close to all members of cluster
  - Average
    - Joins if close to average of cluster members
  - Ward’s Method
    - Based on average sum of squared distances
      - Most ‘statistical’
      - Most stable when further data available

5

- Implementation in **R**
  - Step 2 distance matrix from data collected in step 1
    - Function **`dist(…)`**
    - Default **`t is euclidean`**
  - Step 3
    - Function **`hclust(.,.,.,…)`**
    - Default is Ward’s method
  - Step 4
    - Decide on number of clusters
    - Function **`plclust(.,.,.,…)`** or **`plot(.)`**

6

- Library **`pvcluster`**
- Facilities for calculating p-values
  - By simulation
- More sophisticated alternatives in library **`cluster`**
  - Also many ‘application-orientated’ libraries
  - Functions **`agnes(), clara(), diana(), fanny(), mona()`**
  - agglomerative nesting, large data sets, divisive analysis, monothetic……..

7

- Many good sources
  - **R** help system has examples
  - **TRY IT & SEE WHAT HAPPENS**

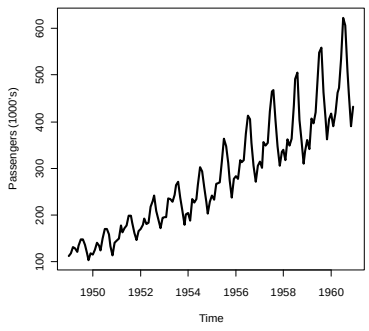
8

- Simple Time series
  - Example: Air Passengers
    - Monthly totals of air passengers

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
1949	112	118	132	129	121	135	148	148	136	119	104	118
1950	115	126	141	135	125	149	170	170	158	133	114	140
1951	145	150	178	163	172	178	199	199	184	162	146	166
1952	171	180	193	181	183	218	230	242	209	191	172	194
1953	196	196	236	235	229	243	264	272	237	211	180	201
1954	204	188	235	227	234	264	302	293	259	229	203	229
1955	242	233	267	269	270	315	364	347	312	274	237	278
1956	284	277	317	313	318	374	413	405	355	306	271	306
1957	315	301	356	348	355	422	465	467	404	347	305	336
1958	340	318	362	348	363	435	491	505	404	359	310	337
1959	360	342	406	396	420	472	548	559	463	407	362	405
1960	417	391	419	461	472	535	622	606	508	461	390	432

9

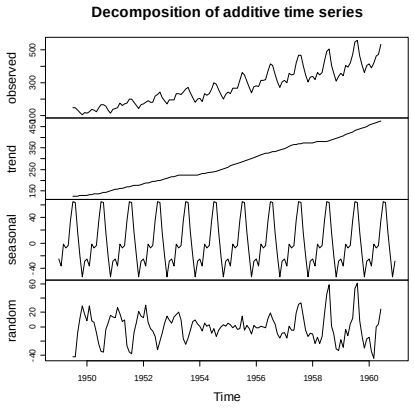
```
ts.plot(AirPassengers,lwd=3,cex=2),)
```



10

- Numbers are generally going up
  - **TREND**
- Annual pattern with peak in mid-year
  - SEASONALITY
- Random
  - Extra 'noise'
- Interest to decompose series into these components

11



12

- Helps choose statistical model for series
  - For understanding 'causes
  - For prediction
    - Predict trend and seasonal components separately
      - (cannot predict random component)
  - May use estimates of random components for matching with other series
    - Cross-matching
    - Cross-dating
    - Tree-rings
    - Pollen sequences
    - ?? Ice cores???