

DESAGREGACIÓN DE BIG DATA GRACIAS A SUS PATRONES DE VARIABILIDAD

Francisco J. Cao García (UCM)

Agregación y Desagregación

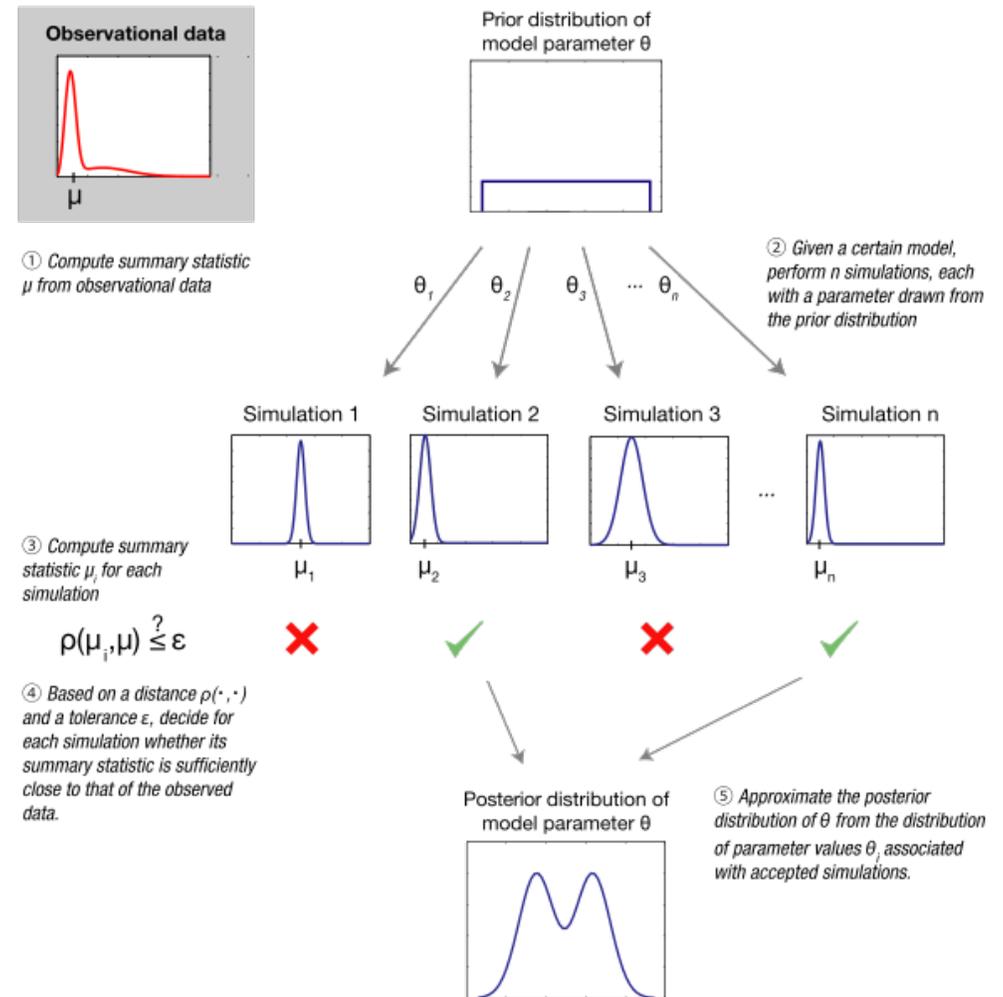
- Agregación:
 - ¿Cuál es el comportamiento habitual?
 - ¿cuantas veces viene el cliente a la tienda? ¿qué productos compra?
 - ¿Cuál es su variabilidad?
 - ¿siempre el mismo día? ¿con el mismo periodo? ¿los mismos productos?
- Desagregación:
 - ¿por qué puede que sea ese el comportamiento habitual? ¿a qué se puede deber?
 - ¿compra aquí todo? ¿solo determinados productos, y otros en la competencia?
 - ¿a qué se deben las variaciones?
 - ¿aprovecha ofertas de la competencia?

Desagregación involucra modelización

- Agregación
 - Cálculo de medias, varianzas, correlaciones, ...
- Desagregación involucra modelización
 - Modelo = descripción simplificada de la realidad
 - Ej.: Los clientes compran cada d días, y hacen la siguiente compra en el mismo supermercado con una probabilidad p , y cambian de supermercado con una probabilidad $1-p$
 - Parámetros = valores que determinan las características de un determinado modelo.
 - Ej.: la periodicidad d , y la fidelidad p en el modelo anterior
 - Las predicciones del modelo han de compararse con los datos y ver si son compatibles

Determinación de parámetros: Inferencia bayesiana

- Escogemos unos valores de los parámetros
 - Ej: $d=7$ (compran cada 7 días), $p=0.7$ (repiten supermercado en un 70% de los casos)
- ¿son las predicciones para esos valores compatibles con los datos observados?
- Así se prueban distintos valores de los parámetros, y se determinan sus valores más probables.



Modelos para secuencias de datos: Hidden Markov Models (HMM)

Si los datos son secuencias temporales

- El modelo consta de estados y probabilidades de cambiar de estados.
- El modelo también incluye para cada estado las posibles observaciones a las que puede dar lugar
- Se sigue el procedimiento indicado de inferencia bayesiana para determinar los valores de los parámetros, que mejor predicen los patrones observados

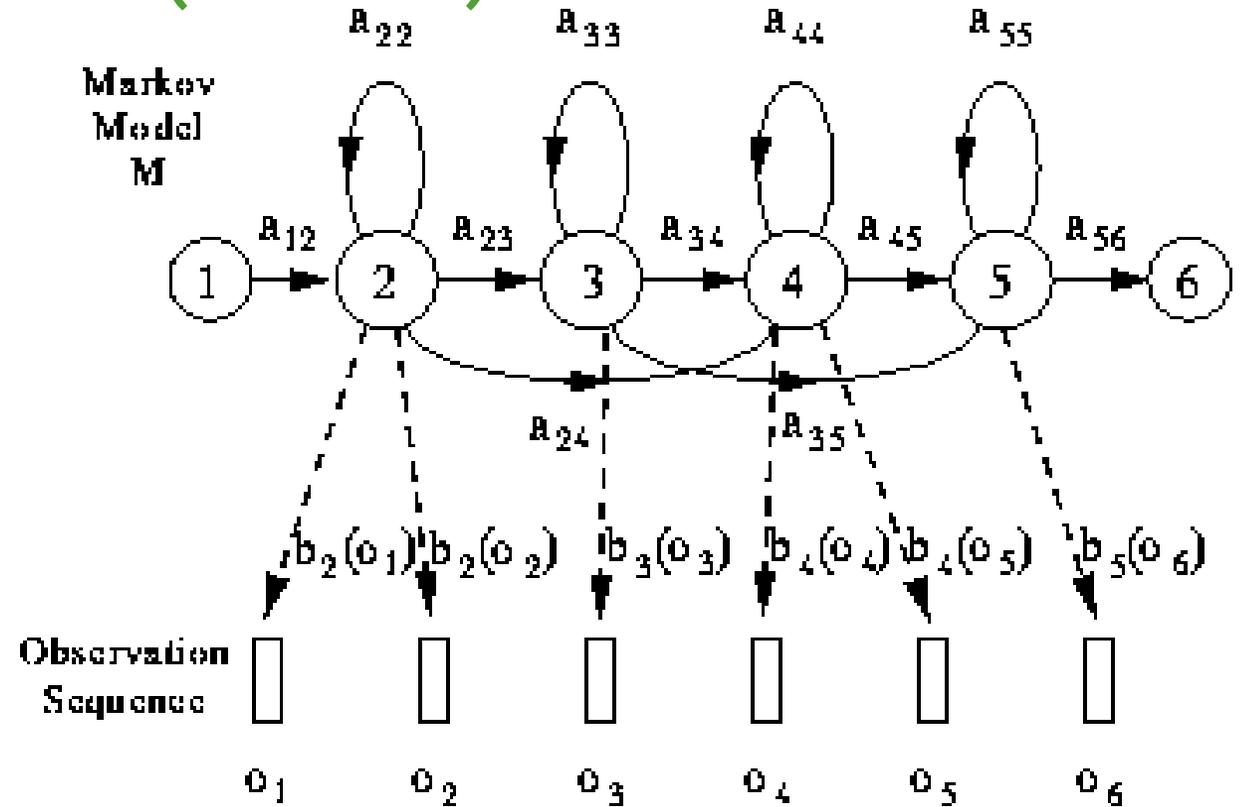


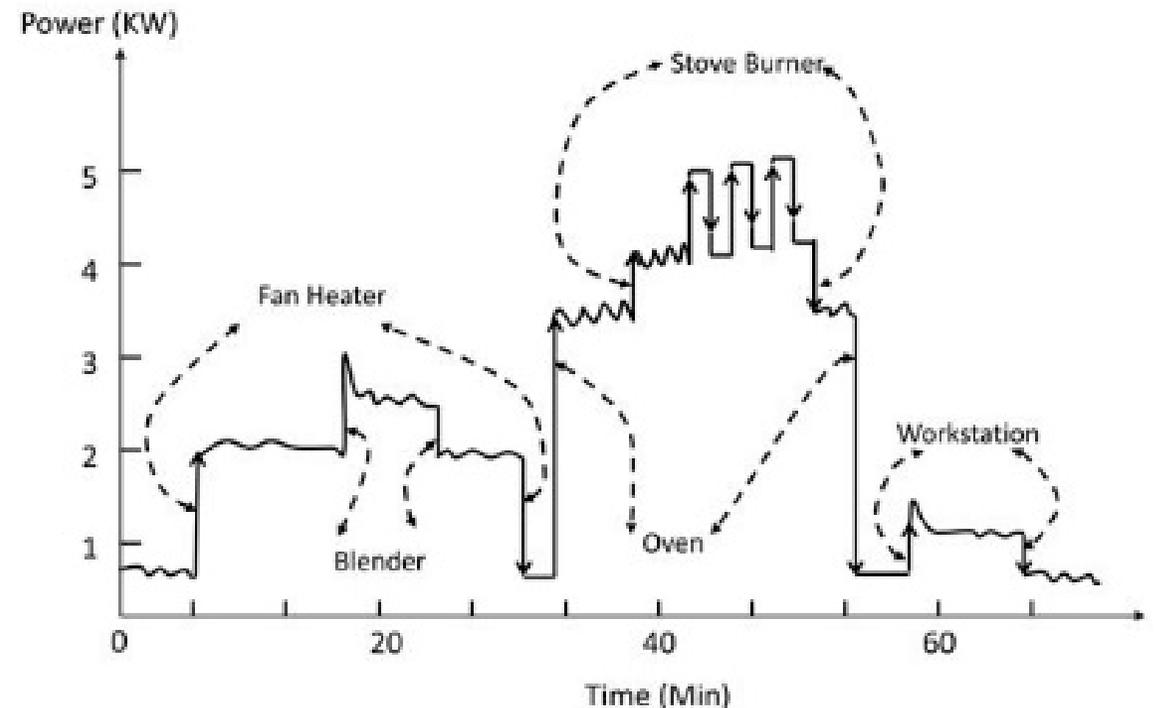
Fig. 1.3 The Markov Generation Model

Aplicaciones destacadas

- Reconocimiento de voz (lo que se ha dicho -> datos: registro sonoro)
- Modelización de diversos procesos biológicos, de los que solo tenemos información parcial. (DNA -> RNA -> datos: abundancia proteínas)
- Desagregación de señales eléctricas (encendido y apagado aparatos -> datos: consumo total)

Desagregación de señales eléctricas

- Identificación de cuales son los aparatos encendidos a partir de la señal de consumo eléctrico.
- Nonintrusive load monitoring (NILM), or nonintrusive appliance load monitoring (NIALM)



Modelos para secuencias de datos: Hidden Markov Models (HMM)

Respuestas:

- Dado un valor de los parámetros probabilidad de obtener los datos observados
- Valores de los parámetros del modelo que hacen más probable los datos observados
- Dada la secuencia de datos observada y los parámetros del modelo, cual es el proceso subyacente más probable

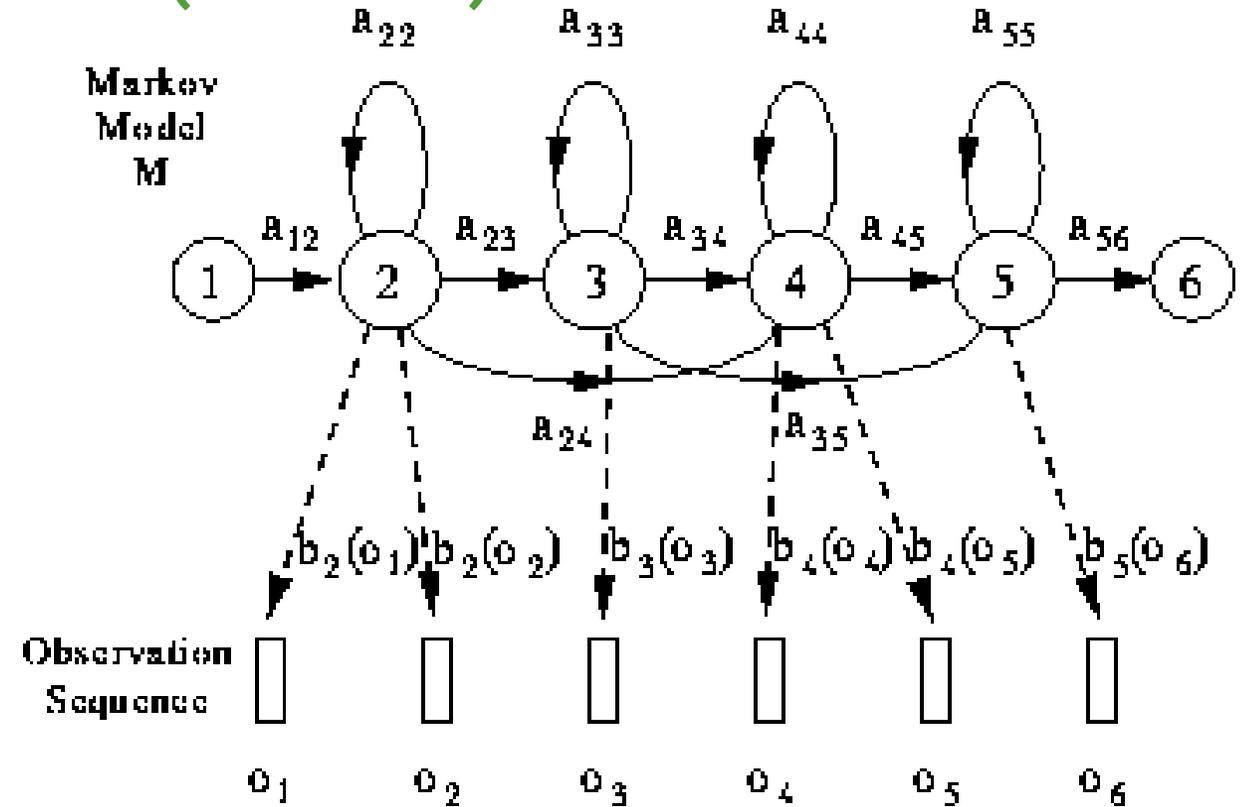


Fig. 1.3 The Markov Generation Model

Conclusiones

- Desagregación abre la posibilidad de obtener más información
- Las cadenas ocultas de Markov (HMM), y sus variantes, permiten desarrollar y determinar los parámetros de modelos que describan los patrones de variación temporal observados.
- Experiencia previa en aplicación de estas técnicas a:
 - Procesamiento de datos biológicos
 - Desagregación de señales eléctricas (para empresas)
- **Oferta para empresas:**
 - **Asesoría en desarrollo de modelos y métodos de determinar los parámetros**
 - **Desarrollo conjunto de modelos y algoritmos.**