FACULTAD DE CIENCIAS ECONÓMICAS Y DE ADMINISTRACIÓN

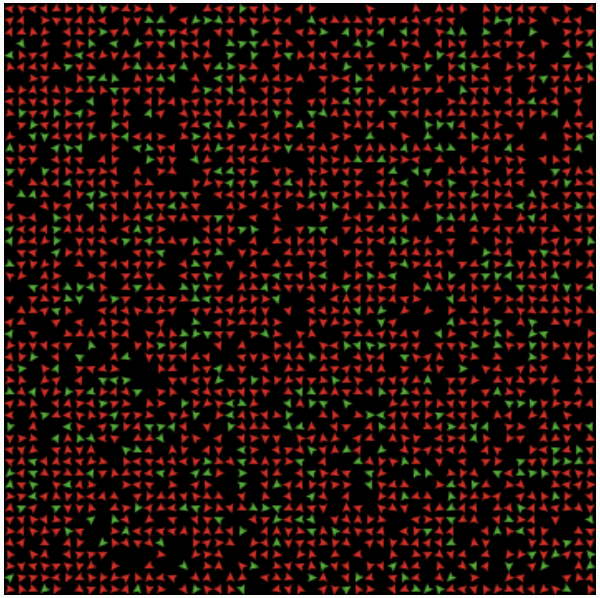
UNIVERSIDAD DE LA REPÚBLICA URUGUAY

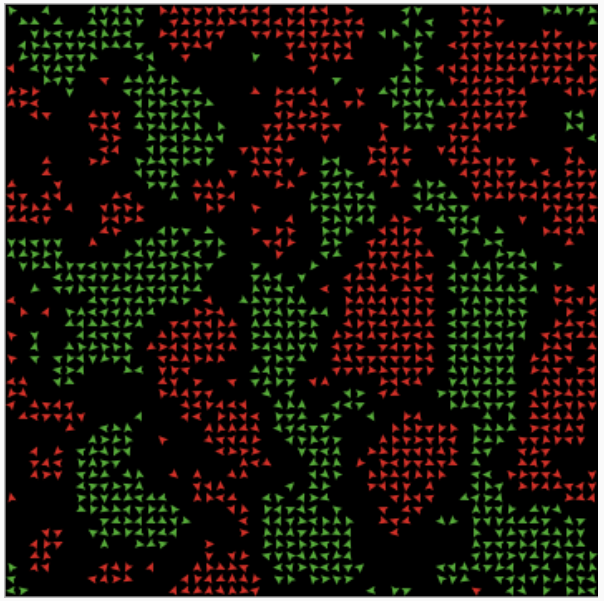
Microsimulación

Docente: Daniel Ciganda

6ta Clase
24 de Septiembre de 2025

Modelo de Segregación de Schelling





- Agentes de dos colores diferentes en un mundo-tablero.
- Cada agente tiene un nivel de tolerancia / umbral que determina su felicidad con respecto a la proporción de agentes en su vecinadrio distintos a si mismo.
- Patrones de segregación emergen en el mediano plazo incluso en presencia de agentes relativamente tolerantes.

Analizando Modelos Complejos

- Nuestros modelos pueden tener una relación compleja y no obvia entre sus **inputs** (preferencia, densidad) y sus **outputs** (segregación).
- Esta relación forma una ”superficie de respuesta” que nos gustaría entender.
- Cada punto en esa superficie requiere ejecutar una simulación que puede ser muy lenta. Explorar miles de combinaciones es computacionalmente muy costoso o imposible.

La pregunta: ¿Cómo podemos entender esta superficie sin tener que simular cada punto?

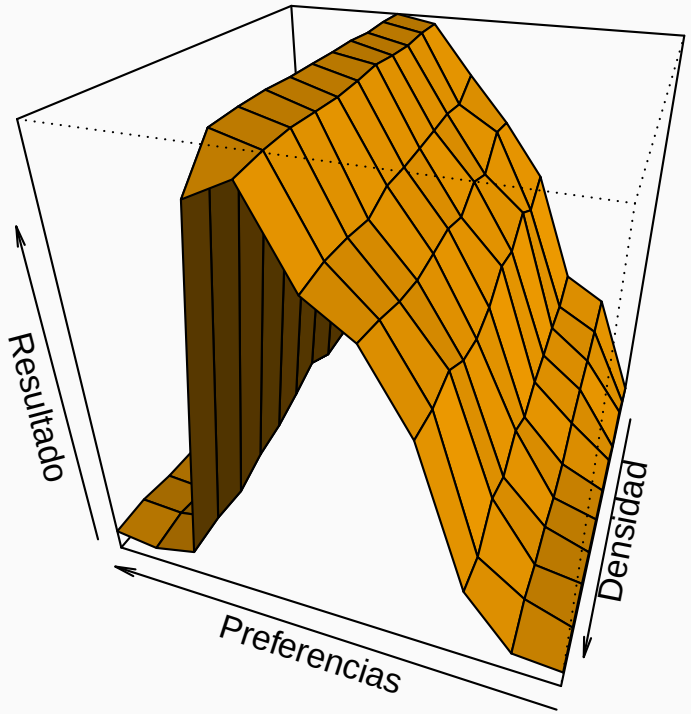


Figure 1: Relación inputs-output Modelo de Schelling

La Estrategia: Construir un Meta-Modelo

- Queremos aproximar una relación desconocida: **outputs** $= f($**inputs**$)$, donde **inputs** $\in \mathbb{R}^p$ y p puede ser grande.
- **Espacio de parámetros enorme**: las grillas crecen *exponencialmente* con p ("maldición de la dimensionalidad").
- Los modelos de simulación suelen ser costosos de correr (tiempo/recursos).
- Un **metamodelo** (o emulador) aprende una aproximación de f para:
  - explorar rápido el espacio de parámetros,
  - mapear superficies de respuesta,
  - hacer validación y sensibilidad sin re-simular.

Cómo construir un Metamodelo (Pasos Generales)

1. **Muestrear el espacio de parámetros inputs**: grilla, muestreo aleatorio, Latin Hypercube, etc.
2. **Ejecutar el simulador** en esos puntos y recolectar **outputs** (ideal: replicar si hay estocasticidad).
3. **Particionar** datos en entrenamiento/validación.
4. **Ajustar varios candidatos** (p.ej., lineal, polinomial, GP, GAM, árboles) sobre el set de entrenamiento.
5. **Evaluar** en validación: predicción vs. observado, residuos, métricas (RMSE, R^2).
6. **Seleccionar y diagnosticar** el modelo que mejor balancee sesgo/varianza.
7. **Usar** el metamodelo: mapas de respuesta, sensibilidad, optimización, escenarios.

Validación

- **Estrategia de Validación Cruzada:**
  1. **Dividimos los datos:** Separamos nuestras 121 simulaciones en un conjunto de Entrenamiento (para ajustar el modelo) y uno de Validación (para probarlo).
  2. **Predecimos:** Usamos el modelo entrenado para predecir los resultados del conjunto de validación (que el modelo nunca ha visto).
  3. **Comparamos:** Graficamos el **Resultado Real vs. el Resultado Predicho**.

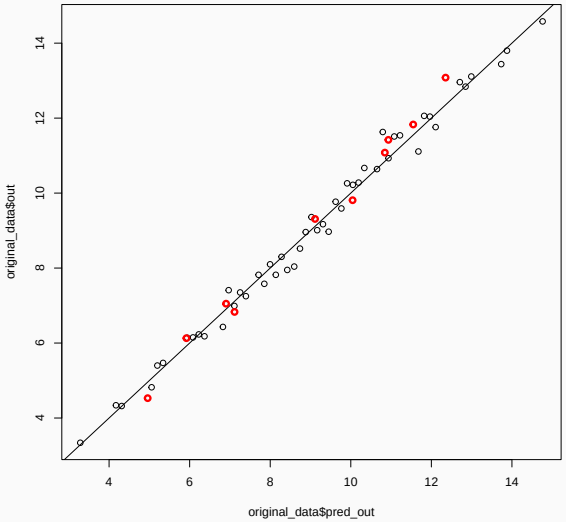


Figure 2: Si el meta-modelo es bueno, los puntos de validación (rojos) deben caer sobre la línea diagonal de predicción perfecta.

¿Qué Hacemos con el Metamodelo?

- **Exploración rápida** del espacio de parámetros (mapas, cortes, escenarios).
- **Análisis de sensibilidad**: importancia relativa de inputs.
- **Optimización/calibración**: buscar regiones que cumplan objetivos (p. ej., minimizar un output).
- **Generalizar**: agregar más inputs, diseños más eficientes, y/o modelos adicionales.

Mensaje final
El metamodelo es una herramienta práctica: acelera preguntas, facilita la validación y hace **explicable** un simulador costoso.

Figure 3: Computación Bayesiana Aproximada

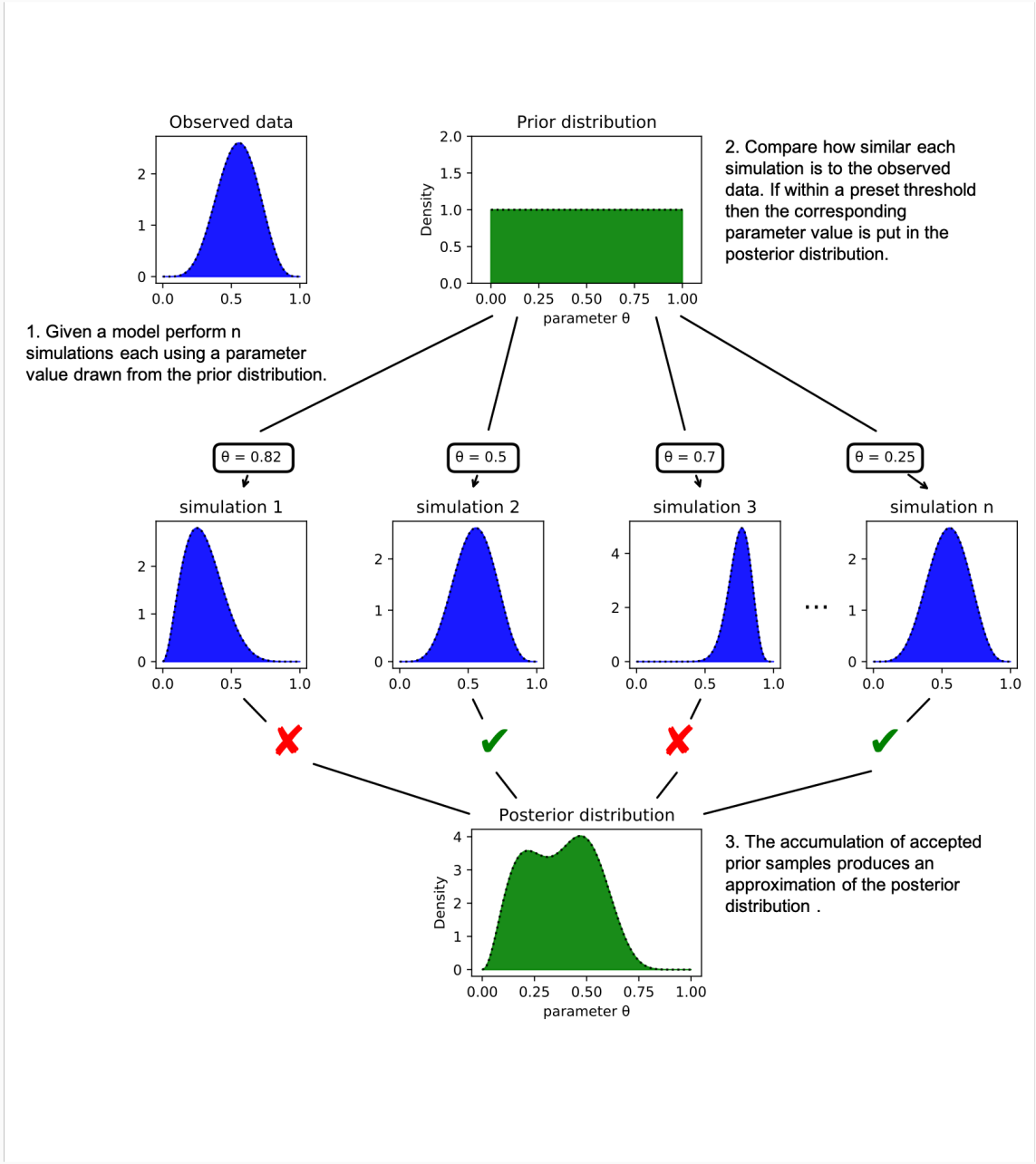
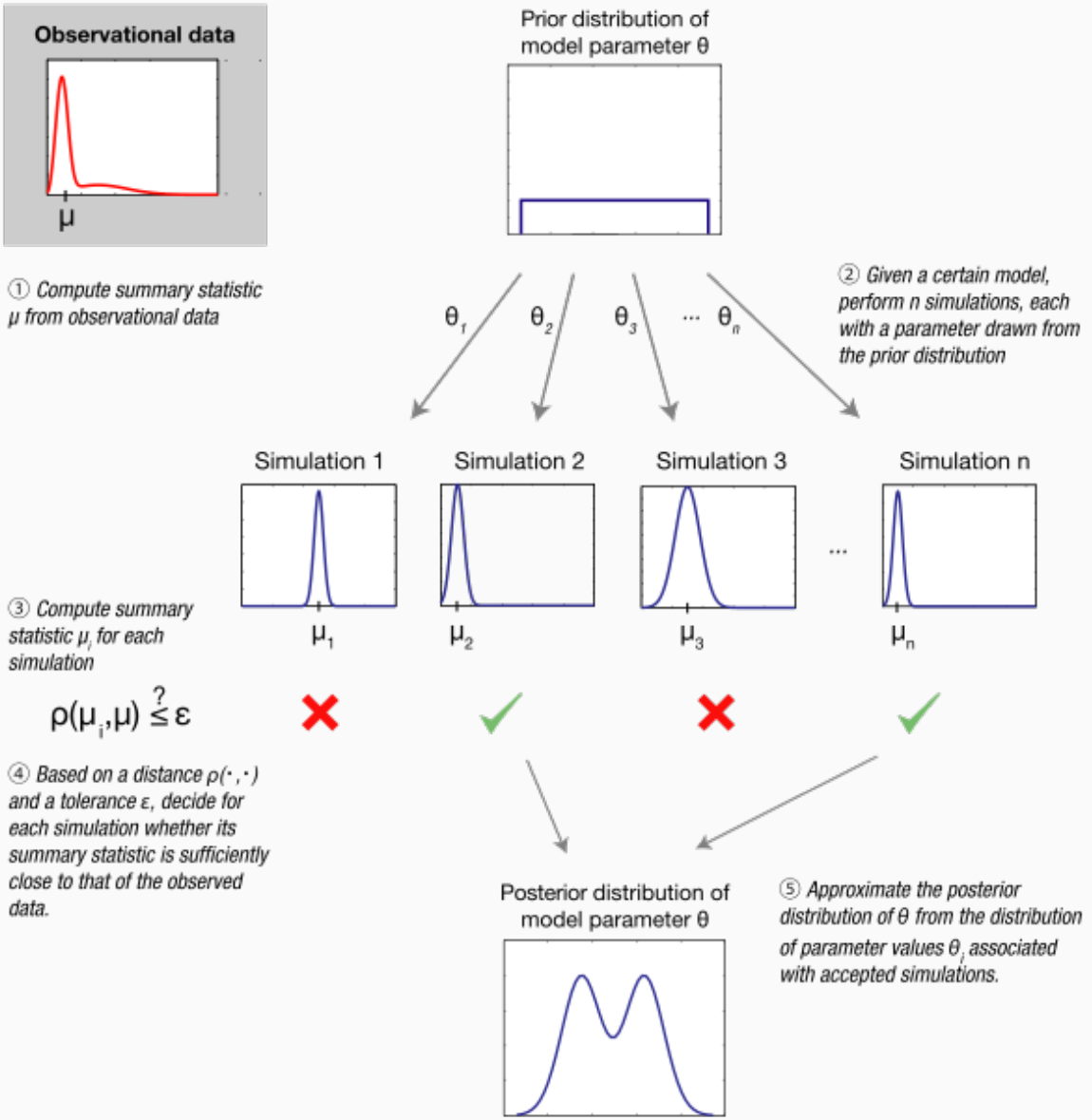


Figure 4: Computación Bayesiana Aproximada



Likelihood-Free Inference

Figure 5: Inferencia Bayesiana

Given a dataset $\mathcal{D} = \{x_1, ..., x_n\}$:

Bayes Rule:

$$P(\theta|\mathcal{D}) = \frac{P(D|\theta)P(\theta)}{P(\mathcal{D})}$$

$P(\mathcal{D}\|\theta)$	Likelihood function of θ
$P(\theta)$	Prior probability of θ
$P(\theta\|\mathcal{D})$	Posterior distribution over θ

Likelihood Free:

- Imposible evaluar la función de verosimilitud
- Posible generar (simular) datos a partir del modelo

Suppose we first draw equally likely values of θ from $p(\theta)$.. For each θ_j, we now draw a D from $f(D|\theta = \theta_j)$ ... Suppose we collect together all D_j that match the observed D_{obs}, and then all θ_j that correspond to these X_j. Formally, this collection of θ values represents the posterior distribution of θ. Rubin (1984)

Inferencia Basada en Simulaciones: ABC

- En la inferencia Bayesiana, todo se basa en el Teorema de Bayes:

$$P(\theta|D) \propto \underbrace{P(D|\theta)}_{\text{Verosimilitud}} \cdot \underbrace{P(\theta)}_{\text{Prior}}$$

- La función de **verosimilitud (likelihood)**, $P(D|\theta)$, es el motor que nos permite aprender de los datos.
- **El problema:** Cuando trabajamos con modelos computacionales no tenemos una expresión matemática para la verosimilitud.
- **Consecuencia:** Los métodos estándar que dependen de evaluar la verosimilitud (ej. MCMC) no se pueden aplicar directamente. Necesitamos un nuevo enfoque.

La Intuición de ABC: Si se Parece, Funciona

La idea central de ABC es simple:

Si no podemos calcular la probabilidad de nuestros datos observados bajo un set de parámetros θ, podemos en cambio buscar los parámetros θ que generan datos simulados que se parezcan a nuestros datos observados.

- En otras palabras, reemplazamos la evaluación matemática de la verosimilitud con un proceso de simulación y comparación.
- Los parámetros que producen simulaciones ”buenas” (cercanas a la realidad) son considerados plausibles y forman nuestra distribución posterior.
- Esto nos permite hacer inferencia Bayesiana para *cualquier* modelo del que podamos simular, sin importar su complejidad interna.

Modelo $\rightarrow$ Generador de Datos

El Algoritmo de Rechazo ABC

El algoritmo más simple de ABC sigue un proceso de 4 pasos:

1. **Muestrear:** Se extrae un set de parámetros candidatos, θ^*, de la distribución a priori $P(\theta)$.

2. **Simular:** Se genera un set de datos sintético, D_{sim}, ejecutando el modelo con los parámetros candidatos: $D_{sim} \sim M(\theta^*)$.

3. **Comparar:** Se calcula una ”distancia” ρ entre los datos simulados y los observados, usualmente a través de estadísticos resumen: $\rho(S(D_{sim}), S(D_{obs}))$.

4. **Aceptar/Rechazar:** Si la distancia es menor que una **tolerancia** ϵ, se acepta θ^* como una muestra de la posterior. Si no, se rechaza.

El conjunto de parámetros θ^* aceptados es una muestra de nuestra distribución posterior aproximada.

ABC en la Práctica: Nuestro Laboratorio de Schelling

El laboratorio que realizaremos implementa exactamente este algoritmo:

- **Modelo (M):** La función ‘schelling()‘.
- **Parámetros (θ):** Los argumentos ‘alikePref‘ y ‘density‘.
- **Prior ($P(\theta)$):** La grilla del diseño factorial (‘full_design‘) que explora el espacio de parámetros.
- **Estadístico Resumen Observado ($S(D_{obs})$):** El valor único de segregación que asumimos como real: ‘observed_segregation = 0.75‘.
- **Distancia (ρ):** La diferencia absoluta: ‘abs(sim_out - observed_segregation)‘.
- **Aceptación (ϵ):** En lugar de un ϵ fijo, nuestro criterio es flexible: nos quedamos con el ‘X%‘ de las simulaciones con la menor distancia.

Objetivo del Laboratorio

Utilizar este flujo de trabajo para encontrar la distribución posterior de ‘alikePref‘ y ‘density‘ que es consistente con el nivel de segregación observado.

Desafíos de ABC (1): El Dilema de la Tolerancia (ϵ)

- El parámetro de **tolerancia**, ϵ, define qué tan ”cerca” deben estar los datos simulados de los observados para que aceptemos los parámetros.

- La elección de ϵ implica un trade-off fundamental entre la precisión estadística y el costo computacional.

Si ϵ es GRANDE

- **Pro:** Alta tasa de aceptación. Computacionalmente eficiente.
- **Contra:** La aproximación es pobre. Aceptamos parámetros que no son muy buenos, resultando en una posterior con alto sesgo (bias).

Si ϵ es PEQUEÑO

- **Pro:** La aproximación a la posterior real es mucho mejor (bajo sesgo).
- **Contra:** Tasa de aceptación muy baja. El costo computacional se vuelve prohibitivo (alta varianza).

En la práctica, no hay un valor ”correcto” de ϵ; es una elección crítica del modelador.

Desafíos de ABC (2): La Maldición de la Dimensionalidad

El problema de la dimensionalidad afecta a ABC de dos maneras:

1. **Dimensionalidad de los Parámetros (θ)**
   - A medida que aumenta el número de parámetros a estimar, el volumen del espacio a priori crece exponencialmente.
   - Muestrear aleatoriamente se vuelve ineficiente. Encontrar la pequeña región de parámetros "buenos" es como buscar una aguja en un pajar. La tasa de aceptación se desploma.

2. **Dimensionalidad de los Estadísticos Resumen ($S(D)$)**
   - Idealmente, usamos "estadísticos suficientes" que capturan toda la información relevante de los datos. Para modelos complejos, casi nunca los conocemos.
   - Si usamos *muy pocos* estadísticos, perdemos información crucial.
   - Si usamos *demasiados*, es casi imposible que un vector de estadísticos simulados esté cerca del vector observado en todas las dimensiones a la vez. De nuevo, la tasa de aceptación se desploma.

Resumen y Soluciones Modernas

En resumen, los desafíos clave del ABC de rechazo son:

- La elección de la tolerancia ϵ (trade-off sesgo-costo).
- La elección de los estadísticos resumen (trade-off información vs. dimensionalidad).
- El altísimo costo computacional.

El Camino a Seguir: ABC Moderno

La investigación actual en ABC se enfoca en algoritmos más eficientes que superan estas limitaciones:

- **Sequential Monte Carlo (SMC-ABC):** En lugar de un solo paso, usa una secuencia de tolerancias ϵ decrecientes, adaptando el muestreo hacia las regiones de alta probabilidad.
- **Ajuste por Regresión (ABC-Post-Processing):** Usa modelos de regresión para corregir las muestras aceptadas, permitiendo usar un ϵ más grande pero aun así obteniendo una buena posterior.
- **Machine Learning:** Se utilizan técnicas para aprender los estadísticos resumen de forma automática o para construir emuladores (como los que vimos en el lab) y reemplazar al simulador lento.

Un Desafío Conceptual: La Equifinalidad

- A veces, el problema no está en el método de inferencia (ABC), sino en la naturaleza del sistema que modelamos.
- **Equifinalidad:** Es el principio por el cual, en un sistema complejo, un mismo resultado macroscópico puede ser generado por múltiples combinaciones de parámetros o procesos microscópicos muy diferentes.
- **Implicación para la Inferencia:** Si diferentes sets de parámetros (θ_1, θ_2) producen resultados casi idénticos, nuestra distribución posterior $P(\theta|D)$ no podrá distinguir claramente entre ellos. El modelo se vuelve no identificable a partir de los datos de resultado.

Ejemplo en el Modelo de Schelling

Imaginemos que observamos un nivel de segregación del 80%. Este resultado podría ser causado por:

- Agentes con **baja tolerancia** en una ciudad de **alta densidad**.
- O por agentes con **alta tolerancia** en una ciudad de **baja densidad**.

Si ambos escenarios son plausibles, nuestra posterior mostrará una correlación entre los parámetros en lugar de un único punto de máxima