

# Combinación de fuentes a nivel intermedio con Stata

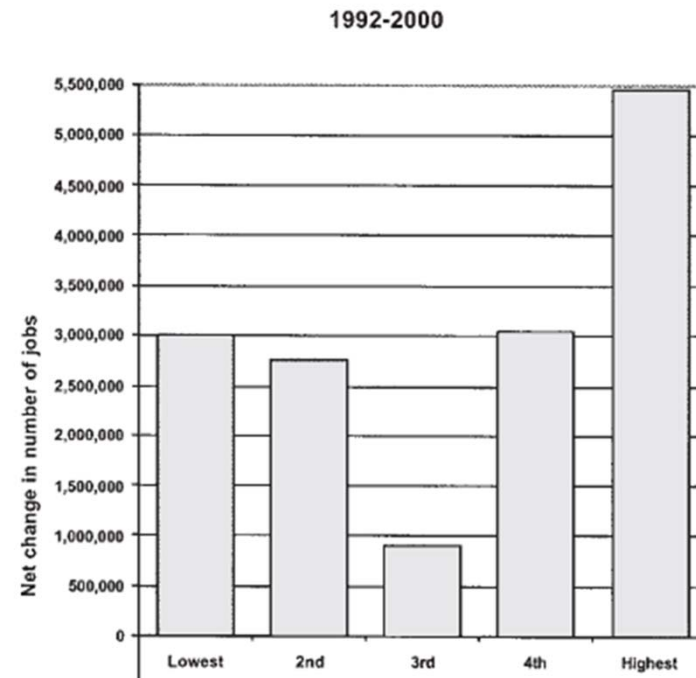
Enrique Fernández Macías

Departamento de Sociología

Universidad de Salamanca

# 1. El objetivo de investigación

- Evaluar si los países de la UE crean “más y mejores empleos”, replicando un estudio que realizaron Wright y Dwyer para EEUU (“the American Jobs Machine”).
- Distribución de la creación de empleo en quintilas de calidad, en función de su salario



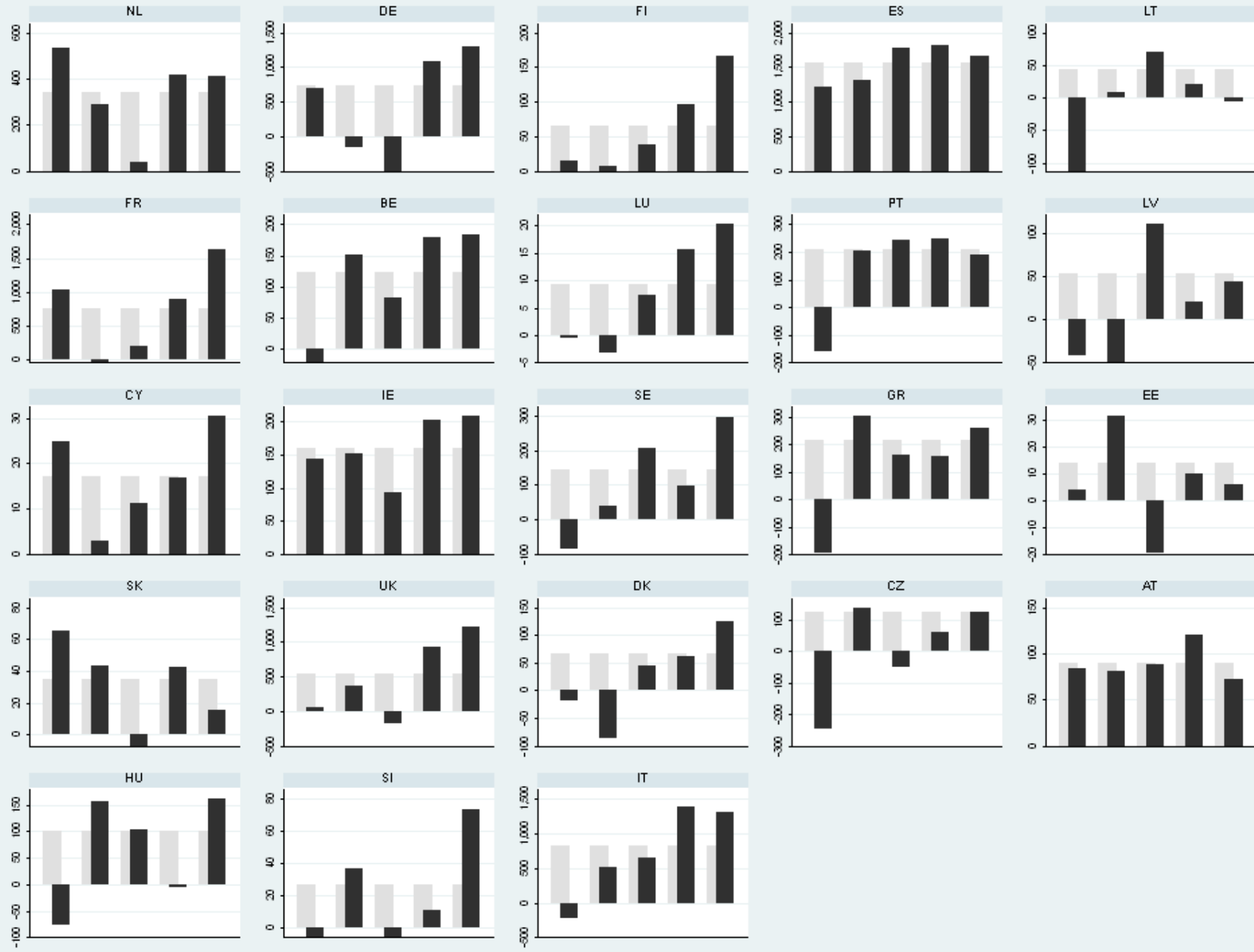
## 2. El problema

- La base de datos utilizada en el proyecto americano (Current Population Survey) incluye salarios y datos de empleo.
- La fuente equivalente en Europa (Labour Force Survey) no incluye datos salariales.
- Distintas fuentes europeas para datos salariales, o bien incompletas (ESES) o bien limitadas en su tamaño muestral (ECHP, SILC).

# 3. La solución

- Aprovechar que el objetivo del proyecto es estudiar el cambio al nivel de sectores y ocupaciones específicos para cambiar la unidad de análisis: empleos en vez de individuos.
- Empleo: cruce de ocupación y sector a dos dígitos (en total, entre 600 y 1500 por país). Variables disponibles en distintas fuentes, y estandarizadas a nivel europeo.
- Construcción de una base de datos nueva, con información agregada de distintas fuentes para cada empleo.
  - LFS para cantidades de empleo, sociodemográficas y laborales.
  - ESES, ECHP, SILC y cuentas nacionales para salarios.
  - EWCS para indicador multidimensional de calidad del empleo

Absolute change (different periods) - wid



## 4. Ventajas e inconvenientes

- + Posibilidad de utilizar muchísimas fuentes
- + Más fácil hacer grandes análisis comparados
- + Mucha mayor riqueza de contenido,  
combinando áreas distintas
- Se pierde variabilidad intragrupal (importante evaluar este efecto *a priori*)
- Menos flexibilidad en el análisis (hay que tener muy claros los objetivos al definir los datos)

# 5. Implementación en Stata

- Estructura básica de la base de datos construida: casos son combinaciones de ocupación y sector a dos dígitos, variables son distintos aspectos de interés.
- En la práctica, para ganar flexibilidad, dos bases de datos distintas, combinadas en el análisis:
  - Datos de empleo, sociodemográficos y laborales: cruce de muchas variables distintas de forma recursiva, con un peso para cada combinación.
  - Datos de salarios e indicador compuesto de calidad del empleo: estadísticos promedio y de dispersión para cada combinación de ocupación y sector.
- Principalmente, son necesarios tres comandos no muy usados pero muy potentes para la manipulación de datos: collapse, reshape y merge.

# collapse (1)

- Sintaxis:

```
collapse (stat) varlist [if] [in] [weight], by(varlist) cw
```

- Admite los estadísticos de resumen habituales
- Admite ponderaciones
- Sustituye los datos en memoria por una base de datos agregados.

- Ejemplo (datos de salarios):

```
collapse (mean) hours (mean) wage (count) cases (rawsum) pop  
[aw=pop], by(country nace2 isco2)
```

Resultado:

country	nace2	isco2	hours	wage	cases	pop
1	1	1	38.78	443.34	37	84.35
1	1	2	40.64	592.96	23	45.43



## collapse (2)

- Ejemplo (datos de empleo): generamos una enorme tabla de frecuencias para cada combinación de variables independientes.

```
collapse (sum) pop, by(country year nace2  
isco2 sex age [etc])
```

Resultado:

country	year	nace2	isco2	Sex	age	...	pop
1	1995	1	1	1	1	...	84.35
1	1995	1	2	1	1	...	45.43

- Ambas bases de datos se manipulan y fusionan para el análisis

# reshape

- Sintaxis:

```
reshape wide/long var1, i(var2) j(var3)
```

- var1: la(s) que define el contenido
- var2: la(s) que define los “casos”
- var3: la(s) que pasa de filas a variables o viceversa

- Por ejemplo, para pasar los años de filas a columnas:

```
reshape wide pop, i(country nace2 isco2) j(year)
```

Resultado:

country	nace2	isco2	pop1995	pop1996	pop1997	pop1998	...
1	1	1	84.35	85.65	89.54	88.44	...

- Para este tipo de análisis, se está continuamente reestructurando los datos con collapse y reshape

# merge

- Sintaxis:

```
merge x:x vars using filename [, opciones]
```

- Se fusionan los datos abiertos en Stata con los datos especificados tras “using”.
- vars especifica las variables cuyos valores se emparejan
- x:x define el modo de fusión: 1:1 cuando los casos son únicos e idénticos para las variables de fusión; m:1 ó 1:m cuando en un fichero son únicos y en otro se repiten; m:m si se repiten en ambos.

- Para fusionar datos de empleo y de salarios:

Si el año está en formato largo, sería:

```
merge m:1 country nace1 isco1 using wages.dta
```

Si el año está en formato ancho, sería:

```
merge 1:1 country nace1 isco1 using wages.dta
```

**Muchas gracias**

efm@usal.es