LMDG Dataset Documentation

Dale T. Mortensen Centre, Aarhus University — Version 15/05/2025

Henning Bunzel

1 ESR_SE_CVR_EVENTS, Relations between *SENR* and *CVRNR*, events data

DSECONAU Updated: 03-04-2022 Special ESR version to ECONAU. *AJOUR_DATO* is 03-04-2022. Last full year is 2021.
ECONAU Updated: 03-04-2022 Special ESR version to ECONAU, last year 2021. Delivered to LMDG project 702728.

1.1 Description

ESR datasets The *ESR* datasets is a special ESR sample made for *LMDG*. It was supposed to be event datasets, but it is a mixture of yearly and event datasets. The yearly datasets are constructed from different versions of *ESR* (*AJOUR_DATO*) made for statistical datasets. The design of the *ESR* database and the creation of versions are not documented. The current version has the latest *AJOUR_DATO* as 03-04-2022.

The following datasets contain the core units, see table 1.

The following datasets define relations between the core units, see table 2.

The following ECONAU datasets contain supplementary variables which can be merged on the relevant core dataset, see table 3.

Dataset ESR_SE_CVR_EVENTS The dataset used for input is *702728\RAW_DATA \VIEWS \ESR_UDV_CVR_SE_TOTAL* which is an event dataset created 03-04-2022

Department of Economics and Business Economics, Aarhus University, email hbunzel@econ.au.dk

Table 1 Core Units

ECON name	LMDG name	Contents	Type	Spell Period	Remarks
ESR_UDV_ARB	*ESR_ARB*	Workplace	Yearly	[2000-2020]	Earliest spell date is 01-01-1983
ESR_UDV_JUR	*ESR_JUR*	Early reporting units	Yearly		
ESR_UDV_CVR_SE_TOTAL	*ESR_SE_CVR_EVENTS*	Firm and reporting units	Events	[2000-1921]	Earliest SE spell date is 01-01-1900. Earliest Rel spell date is 24-11-1848
	ESR_SE_CVR	Relation between SENR and CVRNR	Yearly	[2000-1921]	Created from *ESR_SE_CVR_EVENTS*
ESR_UDV_OK	*ESR_OK*	Economic units	Yearly		
ESR_UDV_KONKURS_TOTAL		Economic units, bankruptsy			

Table 2 Relations between Core units

ECON name	LMDG name	Contents	Type	Spell Period	Remarks
ESR_UDV_ARB_P	*ESR_ARB_P*	Relation Workplace-Production units	Yearly	[2000-2020]	
ESR_UDV_OK_ARB	*ESR_OK_ARB*	Relation Economic-Workplace units	Yearly	[2000-2020]	
ESR_UDV_OK_CVR	*ESR_OK_CVR*	Relation Economic-Legal units	Yearly	[2000-2021]	

Table 3 Supplementary variables, Core units

ECON name	LMDG name	Contents	Variables	Type	Spell Period	Remarks
ESR_UDV_ARB_ADRESSE						
ESR_UDV_ARB_BRANCHE						
ESR_UDV_ARB_P						
ESR_UDV_CVR_ADRESSE						
ESR_UDV_CVR_BRANCHE						
ESR_UDV_CVR_VIRK_FORM						
ESR_UDV_OK_ADRESSE						
ESR_UDV_OK_VIRK_FORM						

as an extraction from *ESR* defining *SENR* spells for (*START_DATO*,*SLUT_DATO*), and a relation between *SENR* and *CVRNR* with the start and end date for the Rel Spells (*REL_START_DATO*, *REL_SLUT_DATO*).

The key for a *SENR* spell is (*SENR*, *START_DATO*).

A Rel spell maps a *SENR* to zero, one, or more *CVRNR* for a specific period and the key is (*SENR*, *CVRNR*, *REL_START_DATO*).

A SENR spell and a Rel spell is still active if end date is equal to '31DEC9999'd.

Sequence for the jobs in esr_se_cvr_events_2022

- The job *_1_esr_se_cvr_events* in directory *esr_se_cvr_events*_2022\ *dan*.
  - It changes still-active-date to 31DEC9999.
  - It selects the date part of the date-time value in variable with type date.
  - It determines the range of dates.
  - It renames a number of variables.
  - It has 2,297,080 observations.
- The job *_2_esr_se_cvr_update in subdirectory dan* updates the *ESR_SE_CVR_EVENTS*.

  - It changes illegal *CVRNR* (cryptized cvrnr blank values) to " ". All observations have a *CVRNR* values, but 323,722 observations have a value '******9509' which most likely is a cryptized " " value. They are changed to " ", see the observations in dataset *data.ILLEGALE_CVRNR*.

- *_3_esr_se_cvr_events_cvr_miss*. Some observations with a missing *CVRNR* have a following spell with a *CVRNR*, see datasets *data.SENR_MISSING_CVRNR_x*. These 2 spells are concatenated with this *CVRNR*. For these observations EDIT8=1. The dataset now has 2,270,249 observations.
- The job *_4_standard_miss_program_se_cvr_events* in subdirectory *dan*. Check the result for missing values.
- The job *_1_esr_se_cvr_events_old_new* in subdirectory *test* compare 2018 dataset ved 2022 dataset. It is not run for this update because the structure is different and th quality is better
- The job *_2_dups_senr_rel_cvrnr_stat_dato* in subdirectory *test* test for duplicates of keys.
- The job *_3_rel_spell_not_in_se_spell* in subdirectory *test*.
- The job *_4_spell_datoer_senere_end_ajour_dato* in subdirectory *test*.
- The *job _5_esr_se_cvr_events_year* in subdirectory *dan* creates the datasets *DATA\YEARS* and *DATA\YEARS_TRUNC* with yearly start and end dates. It is assumed that a Rel spell must be contained with the SE spell. For many observations this is not the case and the job deletes and edit observations observations to make the dates consistent. See the tables **??** and Table **??**.
- The job *_6_se_cvr_events_drop_edit* in subdirectory *dan* drops the variable edit8 used in previous job.
- The job *_5_se_cvr_events_year_overlaps* in subdirectory *test* uses the dataset *YEARS* dataset to test if Rel spells are contained within SE spells and if Rel spells overlaps, and if the key (*SENR*, *CVRNR*,*REL_START_AAR*) is unique.
- The job *desc_esr_se_cvr_events* in subdirectory *desc* provides descriptions of the dataset.

For some observations *START_DATO* is missing. Observations with missing start date all have *SLUT_DATO* before 31-12-1999. Some have a *CVRNR* but most have a missing *CVRNR*.

At any point in time a *SENR* can only be assigned to one *CVRNR*, but a *CVRNR* can be assigned many *SENR*. The *SENR* is assigned by the tax or custom authority when the unit is created. An econonomic unit can always change *CVRNR*.

323,722 *SENR* numbers are not assigned a *CVRNR*. It is typically an activity without employment, but in this dataset some *SENR* are missing because they were created before *CVR* was established in 1999. In the latest years about 200-250 observations have missing *CVRNR*. This is less than the number of units without employment, which according to DS is about 8,000. In the years earlier than 1995 the number of missing values for *CVRNR* is 500-15,000.

We use *CVRNR* to match other data sets, such as *FIRM*, *FIRE*, *BFL*, *CONESR* and in these datasets all units have employment.

All Rel spell observations must have valid values for *SENR*, *CVRNR*, *REL_START_DATO*, and *REL_SLUT_DATO* and the dataset *ESR_SE_CVR_EVENTS* has missing *CVRNR* values, hence the Relation Spells data *YEARS* is created by selecting observations with valid values for *CVRNR*.

Remarks on the 2022 dataset *ESR_SE_CVR_EVENTS* .

- The Extraction date has value 03-04-2022.
- After deleting some observations the dataset has 2,270,249 observations.
- 296,891 observations have missing value for *CVRNR*.
- 35,119 observations have missing valuess of *START_DATO* and *SLUT_DATO*. They are distributed over the entire period.
- The key is (*SENR*, *START_DATO*). It has 56,440 duplicates. The *SENR* are assigned more than 1 *CVRNR* over time.
- The dataset has 2,268,657 unique *SENR*.
- The dataset has 1,883,904 unique *CVRNR*.
- 1,883,919 observations have *SENR* equal to *CVRNR*.
- 1,4099,696 *SENR* numbers are still active.
- The range of *START_DATO* is [01-01-1900, 25-03-2022].
- Rel spells have the key (*SENR*, *CVRNR*, *REL_START_DATO*.) The key is unique.
- The range of *SLUT_DATO* for terminated *SENR* spells is [18-1-1901,31-12-2022].
- 2,297,075 observations have a terminated SE spell, but the REL spell is still active.
- 26,767 observations have SE spell start date later than REL spell start date.
- 77 observations have SE spell start date later than REL spell end date.
- 77 observations have *SLUT_DATO* later than extraction date *AJOUR_DATO*. They all have a date later in 2022 and all *REL_SLUT_DATO* are 31-12-9999. This not a problem as the last year used in spells is 2021.

Remarks on the 2022 Rel spell datasets YEARS and YEARS_TRUNC .

- The Rel spell dataset has 1,947,326 observations .
- See table **??** to see number of observations deleted and edited in the creatin of dataset *YEARS* from *ESR_SE_CVR_EVENTS*.
- In the Rel spell dataset all observations have valid values for *SENR*, *CVRNR*, *REL_START_DATO* and *REL_SLUT_DATO*.
- The datasets are truncated to [1980,2021]
- In the *YEAR_TRUNC* dataset the *REL_SLUT_AAR* is set to *SLUT_AAR*.
- The datasets are used to create *ESR_SE_CVR*. The expected number of observations in *YEAR* is 32,346,279, and the expected number of observations in *YEAR_TRUNC* is 21,347,392.
- The key for Rel spells (*SENR*, *CVRNR*, *REL_START_DATO*) is unique.
- The key for Yearly Rel spells (*SENR*, *CVRNR*, *REL_START_AAR*) is unique.
- The key for Yearly Rel spells (*SENR*, *START_AAR*) has 43 duplicates. There are several (*SENR*, *CVRNR*) spells with gaps within a year.
- The Rel set has 1,946,001 unique *SENR*.

- The Rel set has 1,947,326 unique *CVRNR*.
- 1,373,593 Rel spells are active. The all have a closed SE spell. Maybe this is not an error but by design.
- The range of *REL_START_DATO* is [24-11-1848, 01-09
- The range of *REL_SLUT_DATO* for terminated relations is [27-11-2001, 12-08-2021].
- 1,829,739 *CVRNR* have just 1 *SENR* assigned. The highest number of *SENR* assigned to a *CVRNR* is 473, see Table 4
- The Rel spells do not overlap.
- 26,766 observations have *START_DATO* later than *REL_START_DATO*.
- 7,679 observations have *SLUT_DATO* before *REL_SLUT_DATO*.
- 77 observations have Rel spell before SE spell.
- 0 observations have REL spell start date later than *AJOUR_DATO*.
- The above 3 problems with Rel spells should not effect the match of yearly economic datasets with Rel spell, but there will be Rel spells which maybe cannnot be matched to economic datasets

Table 4 Frequency of *CVRNR* assigned 1,2,..*SENR*

Number of SENR assigned	Frequency of CVRNR
1	1,829,739
2	87,370
3	21,261
4	6,396
5	2,625
< 5	325,483

Table 5 Frequency of *SENR* assigned 1,2,..*CVRNR*

Number of CVRNR assigned	Frequency of SENR
1	2,267,301
2	2,294
3	567
4	52
5	35

The LMDG dataset has been modified:

1. In the raw 2022 dataset all dates are day-time. They are all converted to day format yymmdd10.
2. 323,722 observations have the *CVRNR* value ”xxxxxxx509” replaced with a missing value.

1.2 Remarks on Update

Errors and Questions, Update of LMDG version

- In the new dataset all dates are day-time. They are all converted to day format yymmdd10.
- In 2018 data a new *CVRNR* xxxxxxx509 is associated with 771,898 *SENR*. In 2022 data 323,722 observations have this *CVRNR*. They are changed to ” ”. This *CVRNR* is NOT a valid *CVRNR*, but most likely a wrongly encrypted blank or illegal value. It has also showed up as a new *CVRNR* value in the current *RAS* version! It is important to get this error fixed.
- *DST* has informed that many *SENR* do not have start- and end-date. These existed before *CVR* was created in 1999. It has not been possible to obtain these dates. If the *SENR* has had the same owner then it is possible to assign a *CVRNR*. 200-250 current *SENR* are *VAT* units (foreign companies or Danish units without employment) Should it not be 8,000?.
- In the 2022 dataset a *SENR* spell is added. We assume that the Rel spell must be contained in the SENR spell. This is not the case, see above.
- All 31DEC4712 values for SLUT_DATO are changed to 31DEC9999.
- Rel spell dates have no missing values. 52702 SE spell dates have missing values. The rel start of these observations are distributed over all years, as late as 2022.
- Many observations with CVRNR missing still have rel spell dates.
- The key (*SENR*, *CVRNR*, *REL_START_DATO*) is unique.
- 26,766 observations have *START_DATO* later than *REL_START_DATO*.
- 7,679 observations have *SLUT_DATO* before *REL_SLUT_DATO*.
- 77 observations have Rel spell before SE spell.
- 1,375,593 observations have a closed SE spell but Rel spell is still active. Maybe this is not an error but by design.