European
Commission

$\sigma^2 = \frac{1}{n}\Sigma(x_i - \bar{x})^2$
$S_x^2 = \frac{1}{n-1}\Sigma(x_i -$
$\sigma = \sqrt{\frac{1}{n}\Sigma(x_i - \bar{x})^2}$

Advanced counterfactual evaluation methods

Guidance document

Social Europe

Advanced counterfactual evaluation methods

Guidance document

European Commission
Directorate-General for Employment Social Affairs and Inclusion
Directorate G

This guidance material has been produced based on the work by the following experts: Claudio Deiana, Andrea Geraci and Elena Claudia Meroni (European Commission, Joint Research Centre, Centre for Research on Impact Evaluation)

LEGAL NOTICE

Manuscript completed in October 2019

Neither the European Commission nor any person acting on behalf of the European Commission is responsible for the use that might be made of the following information. More information on the European Union is available on the Internet (http://www.europa.eu).

Luxembourg: Publications Office of the European Union, 2019

PDF ISBN 978-92-76-11881-7 doi:10.2767/464242 KE-04-19-627-EN-N

Purpose of the guidance document

The aim of this guidance material is to provide an accessible introduction to advanced counterfactual impact evaluation (CIE) methods. The guidance material is mainly intended for the managing authorities of the European Social Fund (ESF) and European Social Fund Plus (ESF+). It can also be used by other institutions involved in implementing and evaluating operations funded by the ESF and ESF+ and by the wider evaluation community.

Focusing on results is one of the main principles of EU investments, and quality evaluations are essential in providing credible evidence about the performance of EU-funded interventions. CIE methods help to answer the fundamental question of whether or not the intervention actually worked. The European Commission encourages Member States to carry out CIEs in order to have more credible evidence about the impacts of the ESF and ESF+ funded operations. Carrying out a high quality CIE is not easy, and the current guidance material is one element of the broader support the Commission gives to the Member States in conducting high-quality CIEs.

This guidance material builds on the previous publication by the European Commission, Directorate-General for Employment, Social Affairs entitled 'Design and commissioning of counterfactual impact evaluations'[1], which introduced the principled and use of CIEs and explained the most common CIE methods. It also complements the support for CIEs provided by the European Commission, Centre for Research on Impact Evaluation[2], at Joint Research Centre which includes video-tutorials, guidance notes, examples of CIEs and other materials.

[1] https://publications.europa.eu/en/publication-detail/-/publication/f879a9c1-4e50-4a7b-954c-9a88d1be369c/language-en

[2] https://crie.jrc.ec.europa.eu/

1. Introduction

The guidance note describes some recent developments of counterfactual impact evaluation methods. The three methodologies proposed are sequence analysis, dynamic matching and synthetic controls.

Sequence analysis can be used in combination with other matching techniques, such as propensity score matching, when there is information on some sequences of an event happening before the treatment. In the case of active labour market policies, knowing the sequence of events and employment statuses that occur before the policy starts can be relevant information to be used in combination with matching on other relevant observable characteristics.

Dynamic matching is also useful when evaluating active labour market policies, in particular in the case of ongoing programmes, i.e. programmes which can start at any point in time and where eventually all the individuals can be treated at one point.

Synthetic control is an extension of difference-in-difference, and can be used to evaluate a policy which affects a whole single region or country, using a pool of non-treated regions or countries as a comparison group.

Details of these three methods are reported in the three sections below.

2. SEQUENCE ANALYSIS

2.1. Background

Sequence analysis can be described as a set of statistical techniques applied within the stream of the sociological literature dealing with life-course studies. In this context, sequence analysis has been developed as a **method of empirically describing and investigating properties, outcomes, and determinants of individual trajectories**. The method was introduced in the sociological literature by Andrew Abbott in the mid-1980s, and has been applied since then in studies dealing with different topics, including family formation, health trajectories and career trajectories.

While a more detailed methodological definition of a sequence (trajectory) is given in the next section, it can be broadly described as the representation of a longitudinal process by a series of states in successive time periods (e.g., the succession of states of employment, unemployment, training, and inactivity of an individual in the last 2 years).

In the context of life-course studies, the sequence per se is the object of primary interest. A number of techniques originally developed in non-sociological fields (mainly computer science) have been applied to study the features of longitudinal sequences for the purpose of uncovering patterns of social change, i.e., how trajectories change across different cohorts and countries[3].

An obvious question is why and how these tools, aimed at a quite different exploratory task, can be of help in estimating the impact of policies.

The core idea is that the past matters, and sequence analysis provides some useful tools to condense the information about the past contained in the individual trajectories. If two individuals are observed at a specific point in time, some of their observable characteristics can be measured. Looking at this snapshot, the two individuals might seem alike in terms of some (or all) of these observable characteristics, and standard matching techniques could be applied. However, their respective past might be substantially different. In other words, if we were to observe the sequence of events that led to the instant in time when the snapshot was taken, then it could be possible that these trajectories were very different. Individuals with similar trajectories might be more alike, and this might play a crucial role in matching performance. For instance, in the field of labour economics, work histories could be used as a proxy for individual ability and help mitigate selection issues.

The key contribution made by the tools of sequence analysis is to provide an operational definition of the similarity between two different sequences based on the properties and the ordering of the elements that they are composed of. The advantage for policy evaluation is that these similarity measures can then be used in a matching approach to identify more appropriate controls for units subject to intervention.

2.2. Method and data requirements

The first step of any analysis of sequences is to define what the sequence is. A sequence can be broadly seen as a way of summarising the information contained in longitudinal data. The latter needs to contain enough

[3] Aisenbrey and Fasang (2010)

detail to be able to re-construct the entire sequence of events that took place before the treatment under study. Consequently, data from administrative registers are likely to play a major role in this context.

From a methodological point of view, a sequence can be described as a string containing a finite number of characters, each representing a state where the individual has been in the time-span between the beginning (hereafter t_0) and the end of the observational period (hereafter T). This time span is related to the study and has to be decided beforehand. In the case of applications dealing with policy impacts, this observational period is generally the pre-treatment period, and its end-point corresponds with the beginning of the intervention. For example, consider the simple case of yearly observations with the pre-treatment period of 3 years, where each individual can be only in one state in each year. The states in this example are represented by the letters A, B, C, or D. Alternative sequences could be represented by:

ABA AAB AAA ABC BBD

Each letter in the above sequences stands for a different state of the world in a given period of time (in this case a year). As an example, in the context of career trajectories A, B, C, D could represent different labour market states such as: employment, unemployment, maternity leave and training.

It can be noticed that one of the fundamental characteristics of the sequence is the list of alternative possible states. This is referred to as the 'state-space'. It is a full list of states of the world mutually exclusive in time, in this case A, B, C, D. Identifying this list of possible states depends on the particular situation.

When dealing with register data, observations are rarely available at yearly intervals, as would be the case in standard longitudinal surveys. In these cases the data can be organised in order to be spaced monthly, and monthly records made use of. The ABA sequence above could then become the following sequence (defining 12 states each year, and observing the transition between A and B in January of the second year):

AAAAAAAAAAAA BBBBBBBBAAAA AAAAAAAAAAAA

Once the state-space has been defined and the data coded accordingly, the result is a string for each individual containing the full sequence of events occurring between t_0 and T. The next step is to adopt a method to compare sequences across individuals in order to identify those that are 'more similar'.

The most common method used in sequence analysis to compare sequences was applied in sociology by Andrew Abbott (1995) and is called 'optimal matching'. The core idea of optimal matching is to construct a measure of similarity between two sequences, 'i' and 'j', based on the edit operations required to transform one sequence into the other. There are two types of operations available: substitution and insertion/deletion ('indel': in-insertion, del-deletion).

As an example, consider the instance of the sequence AABC and the sequence ABBD. The first can be transformed into the latter by deleting the first element A (deletion), adding an extra B (insertion), and substituting the final C with a D. Another possibility would be to substitute the second A with a B, and the last C with a D, which is to say that there are a number of possible ways to 'align' two sequences.

The choice between alternative alignments depends on the 'cost' of each operation. The core idea is that each edit operation comes at the cost of altering the original sequence. The issue of the link between costs and operations is widely discussed in the methodological literature. The most common method used in the social sciences is to apply a data-driven approach to set substitution costs, and to avoid using indel operations especially in the analysis of equal-length sequences.

As described by Lesnard (2010), 'When an event is inserted or deleted, it is also time that is either added or removed. On the other hand, substituting an event by another preserves the timing of the sequences, but at the cost of approximating an event by another.' In the example above, after the first deletion of A, the remaining bit of the sequence, ABC, 'slides backward' to match the other one. This implicitly amounts to assuming that the same events (AB) occurring at different times are perfectly identical (zero cost of transformation).

As for the computation of substitution costs, the data-driven approach mentioned above suggests that these costs should be set using the inverse of the transition frequencies between two different states. For example, if the transition between A and B at time t in the data at hand happens to have a probability of 0.1, then substituting A for B would "cost" 1/01=10[4]. It should be noted that what is important is not the cost per se, but its relative value. Intuitively, it is desirable that transitions that occur more frequently are associated with lower costs of substitution.

When sequences are of unequal length, for example, due to the existence of gaps, indel operations are indeed necessary. In these cases the rule of thumb identified in the literature is to set the cost of indel operations equal to half the maximum substitution cost.

Finally, the distance between two sequences, — 'i' and 'j' - is computed as the sum of the costs implied by the transformation of 'i' into 'j'.

The final question to be answered is how this procedure helps in identifying an appropriate control group for units subject to an intervention. The key principle is that low transformation costs between sequences mean a short distance between them, which translates into similarity between individuals.

Imagine having a list of units in the treatment and control groups, and sequences representing the succession of states in the pre-intervention period. The distance between the sequence of each unit in the treatment group and each of the sequences of units in the control group can be computed. Once the distances have been computed using optimal matching, the closest control unit(s) represent the most similar individuals, and so can be chosen as a match for the treated unit.

As in the simpler case of propensity score matching or exact matching, the *average treatment effect on the treated* is ultimately obtained as the difference between the average value of the outcome for the treated units and that of the matched control units. The assumptions that allow this causal effect to be identified are the same as those needed in the case of propensity score matching and exact matching. Significantly, only the selection of observable characteristics can be accounted for, so any selection bias arising from differences in unobservable characteristics between treated and control units may affect estimates of the causal effect. However, this method enriches the information set available by exploiting the longitudinal dimension of the

[4] The transition probability is computed as the fraction of individuals transiting from state A to state B at time t out of the total number of individual in state A at time t. The full set of all possible transitions can be summarised in a transition matrix, and can be computed separately for each time period (e.g., each month).

data. Adding the full trajectory observed before treatment in order to match treated and control units should help to mitigate any concerns about the selection bias, as this makes it possible to better control for the role of unobservable characteristics. For example, when interventions targeting unemployed individuals are evaluated, observing a worker's past employment history can add a lot of precious information that can be used in the matching. By considering the sequences of events occurring before a given intervention, it is possible to match treated individuals with non-treated individuals similar not only in terms of the usual observable variables normally available (gender, level of education, age, etc.) but also in terms of their labour market history, which, as shown by Caliendo et al. (2017), is as good as controlling for personality traits (which are normally not available for the evaluators in the set of observable characteristics).

Longitudinal data from administrative registers are the best candidate for applying the method described when evaluating policy. This type of data typically enables tracking individuals through time, and contains enough information to re-construct pre-treatment trajectories after the state-space has been carefully defined.

2.3. How it works in practice: An example from the academic literature

Causal Effects of the Timing of Life-course Events: Age at Retirement and Subsequent Health. (Barban, N., De Luna, X., Lundholm, E., Svensson, I., & Billari, F. C. 2017)). *Sociological Methods & Research*

The work by Barban et al. (2017) is a perfect example of how sequence analysis techniques, and optimal matching specifically, can be applied to causal inference. The authors wanted to evaluate the effects of the timing of retirement on subsequent health. Retirement is the treatment individuals are exposed to: individuals self-select into treatment, and this may cause selection bias. More specifically, individuals retiring earlier on a voluntary basis are likely to have worse health conditions as well as worse health future outcomes compared to those retiring later. The authors wanted to identify an appropriate control group able to represent what would have happened to early retirees had they stayed at work, in order to estimate the causal effect of voluntary early retirement.

The authors applied a standard matching based on propensity scores, using pre-treatment characteristics. The chosen variables used to calculate the propensity score were: education, income, marital status, partner's retirement, unemployment status and health status.

The authors highlight that the full observable health trajectory of an individual is also likely to play a key role in shaping retirement decisions while also being a potential factor influencing post-retirement health. In order to exploit the information contained in the pre-retirement health history, the authors use sequence analysis techniques.

The authors use yearly data from linked longitudinal administrative records from different sources. Since the interest is in matching pre-treatment health trajectories, the state-space is defined by 8 potential states ranging from 'No hospitalization or any health related benefits in year t' to 'Spent more than 3 days in hospital during year t', and encompasses other potential states comprising a mixture of hospitalisation and/or benefits in year t.

Having defined the state-space, the authors match treatment units (early retiree) with their nearest neighbour in the control group by applying the optimal matching algorithm on the individual pre-retirement health trajectories.

The final step also combines propensity score matching based on pre-treatment covariates with optimal matching based on the full health trajectories. It should also be noted that, given the large sample dimension typical of register data, the authors use a combination of propensity score and optimal matching on health trajectories: the authors first match treated and control units on year of birth and educational level at time of retirement, and then they combine the other two matching techniques within each cell.

The graphs in Figure 1 are taken from Barban et al. (2017) and show the evolution of the outcome of interest and the number of hospitalisations before and after early retirement for men. The cut-off for early retirement for men (i.e. the treatment) is set at the age of 61. The comparison between treated and control units is presented separately for each of the matching procedures described above, i.e. propensity score, optimal matching on health trajectories, and a combination of these two.

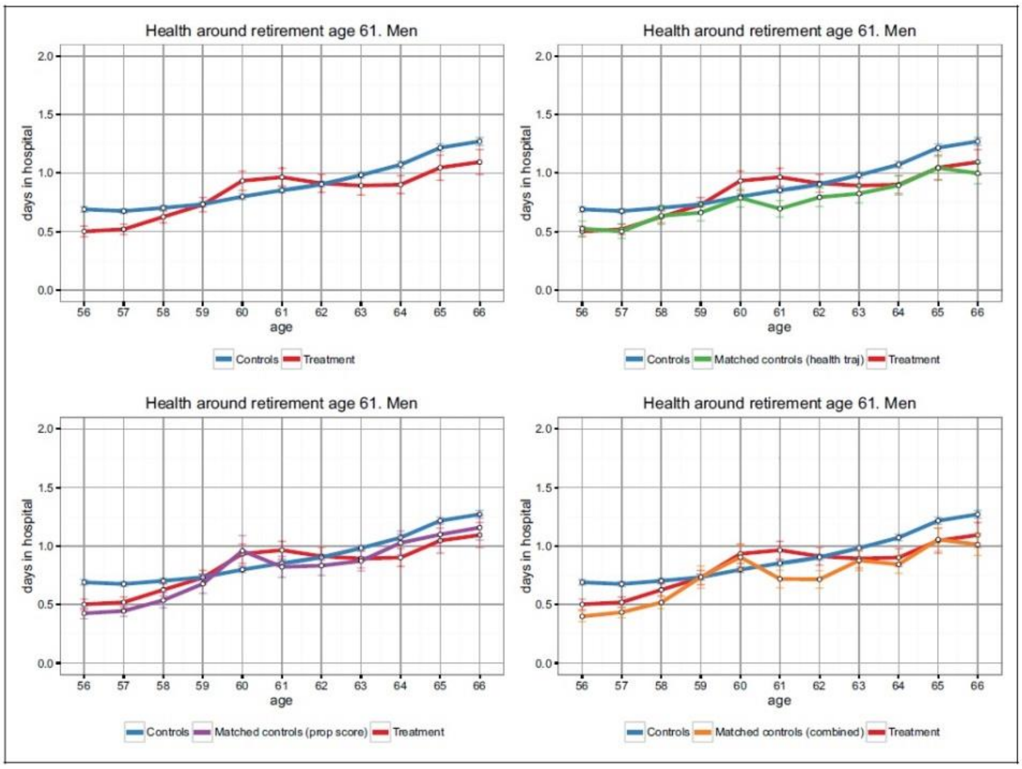

Figure 1 – Source: Barban et al. (2017). The outcome of interest is the number of hospitalisations before and after retirement for men. The threshold for early retirement is set at the age of 61.

One crucial finding in their study is that the pre-retirement health trajectory is a major confounding factor. While it is true that early retirees exhibit faster deterioration of their health, the decision to retire early is also influenced by the full preceding health trajectory. Once the latter is accounted for, the magnitude of the negative effect of retirement on subsequent health (measured as the number of hospitalisations) shrinks substantially.

2.4. References

Abbott, A. (1995). Sequence analysis: new methods for old ideas. *Annual review of sociology*, *21*(1), 93-113.

Aisenbrey, S., & Fasang, A. E. (2010). New life for old ideas: The 'second wave' of sequence analysis bringing the 'course' back into the life course. *Sociological Methods & Research*, *38*(3), 420-462.

Barban, N., De Luna, X., Lundholm, E., Svensson, I., & Billari, F. C. (2017). Causal Effects of the Timing of Life-course Events: Age at Retirement and Subsequent Health. *Sociological Methods & Research*

Caliendo, M., Mahlstedt, R. and Mitnik, O. (2017): Unobservable, but Unimportant? The Relevance of Usually Unobserved Variables for the Evaluation of Labor Market Policies, *Labour Economics*, 46, 14-25.

Lesnard, L. (2010). Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns. *Sociological Methods & Research, 38(3), 389-419*.

Sequence analysis

PROS AND CONS

+ It is a useful tool for condensing the information about the individual's past trajectories.

+ It can be used in combination with standard matching techniques to add information to the usual observable characteristics used when performing matching, which should help to mitigate concerns about the selection bias.

- It is data demanding, as the data used should contain enough information to be able to re-construct pre-treatment trajectories.

3. DYNAMIC MATCHING

3.1. Background

Dynamic matching is an extension of the widely used matching technique called static matching. Static matching is usually applied when a group of individuals receive a treatment (e.g. participation in a programme), and this group is matched to a group of individuals not receiving the treatment according to a set of observable characteristics. The impact of the treatment is recovered by simply comparing the outcome in the treated group with that in the selected similar control group. There are various techniques that can be used to match individuals, the most common being propensity score matching. The main assumption for the estimate to represent the true causal effect of the treatment is that the matching is based on all characteristics that affect selection into the treatment and future outcome, and that there are no unobservable characteristics which influence selection into the treatment[5].

An example when static matching can be used is a programme administered at a fixed point in time, where unemployed individuals are either treated (that is, participate in the programme) or not treated (that is, do not participate). Among the pool of non-treated individuals, a group resembling the treated one in terms of observable characteristics has to be selected. The matching approach assumes that participation in the programme and future labour market status only depends on observable characteristics and there are no unobserved differences in motivation or unobserved ability between treatment and control groups. The effect of the treatment is estimated by comparing the labour market outcome (i.e. whether the individuals are employed or not) of the treated with that of the controls, a reasonable amount of time after the programme has ended.

On the other hand, dynamic matching is useful in the case of ongoing programmes where any eligible individual can potentially become a participant sooner or later as long as they are still eligible, because the programme is offered constantly. Using static matching in this setting can have limitations: if one wishes to evaluate the impact of the programme by comparing participants to similar non-participants by using static matching techniques, a predetermined point in time would need to be set in order to identify those who have actually been treated by that time, and those who still had not been treated. For example, when evaluating a programme for the unemployed, the fixed point could either be an exact date (all those starting the programme by April 20xx are considered to have been treated, while those not starting by that date are placed in the potential control group); alternatively a fixed number of months after registration into unemployment (all those starting the programme within 6 months after the registration are considered to have been treated, while those not starting at all or starting later are placed in the potential control group). However, this approach would be incomplete because the choice of participating in the programme is not made once and for all, as individuals can choose to join it at any time. Therefore, as a comparison to the treated group, a group of individuals who were never treated simply cannot be chosen because they may represent a very different group of individuals. Similarly, it is not appropriate to set a fixed time threshold since the choice that an individual is facing at any

[5] This assumption is called ‘conditional independent assumption’.

given moment for an ongoing programme is not whether to start a programme or not at all, but whether to start it *now* or not to start it *now*.

3.2. Method and data requirements

In dynamic matching the evaluator compares individuals who are similar up to a given point in time, but at that point some receive the treatment and others do not receive it (yet). The treatment is therefore starting a programme (in a given month). What is estimated is the impact of joining a programme at a given time compared to not joining at least up to then. Therefore, the estimated effect is different from the impact of participating in a programme rather than not participating at all, or from the effect of participating in a programme from time *t1* rather than from time *t2.* The comparison group is composed of eligible individuals who have not participated in a programme yet, including individuals who may participate in the future.

The dynamic matching method is a propensity score matching generalisation used to estimate a specific matching estimator for each time interval. This methodology iterates the propensity score matching estimator (or other matching estimators) over different time spans while dynamically (re)defining the control group. This approach requires individuals to be observed through a set of points in time, possibly close to each other, e.g. months, which makes it more data demanding than the standard static framework. In practice, the propensity score is estimated a given number of times corresponding to the number of months (or group of months) of interest. In each period there will be a new group of treated, i.e. those starting the programme in month *u,* and a new group of non-treated, i.e. those still not participating in month *u*. In this process, the treatment is continuously updated vs. the control-group definition over the time span involved. At the beginning of each time window the eligible population comprises individuals who were still not participating in the programme at the end of the preceding period. Individuals starting the programme in month *u* are matched with those never treated or starting the programme at month *u+1* or later. Therefore, the comparison group at *u* is composed of all those still not engaged in the programme at *u*, irrespective of what happens after *u*. Some of them may go into a programme later whereas others may never participate. For both controls and treated at *u*, whatever happens after *u* is viewed as an outcome of joining a programme vs not joining at *u*.

The evaluators can then compare these estimates and find an answer to questions like ‘When is it better to start a specific programme for the eligible population?’ and ‘Is a programme starting at a given point in time still effective for the eligible population?’. In addition, the average of the effects through time can be computed in order to obtain a general answer about the effectiveness of the programme overall.

The causal effect of interest is identified under two main conditions. First, just as in the standard static framework, dynamic matching is based on all characteristics that affect selection into the treatment and future outcome, and on there being no unobservable characteristics which influence selection into the treatment (the conditional independence assumption), which must hold sequentially. Just as for the static matching, this assumption requires detailed knowledge of the factors that drive participation as well as availability of data suitable for capturing those participation determinants that are also likely to affect outcomes. In addition, in the dynamic setting, the evaluator needs to observe monthly or quarterly information about the treated and non-treated population, and the sample size should be large enough to perform reasonable matching in each of these months/quarters. Consequently, administrative data containing rich observable variables capturing the background of individuals are the most appropriate data for applying this methodology. Secondly, identification requires that, conditional on observed covariates, the exact timing of current and future treatment statuses and future outcomes cannot be anticipated by individuals (no anticipation assumption). Examples of when violation of this condition would happen are: 1) if an individual refuses to participate in a programme because he knows that he will be offered an opportunity to participate in another programme in the next few months – *anticipation of future treatment status;* 2) if an individual refuses to participate in a programme because he has already received a job offer which will start in the near future (e.g. seasonal workers, who know that they will be called back from their employer) – *anticipation of future outcome status.*

According to the details of the data and the sample size available, different time windows can be chosen: if the data is very detailed and the sample size large enough, individuals can be compared on a monthly basis. In contrast, if sample size is limited or information is only collected quarterly, several months can be grouped together.

The effect of interest can be defined for each time window (whether it is a single month or a group of months) given that the treatment is received within that time window.

In addition to the evaluation of single programmes, dynamic matching can also be used to evaluate systems of active labour market policies, composed by different actions taking place continuously. (See Sianesi, 2004 summarized in the following section). Finally, dynamic matching can be extended to the multiple treatment framework and allows comparisons of the effectiveness of many different programmes. In this case, rather than comparing the individuals participating in a programme to the ones still in open unemployment, one has to compare them to the ones participating in another programme. By doing this, the effectiveness of one programme can be directly assessed against the effectiveness of another. (See Sianesi, 2008)

3.3. How it works in practice: An example from the academic literature

An evaluation of the Swedish system of active labour market programmes in the 1990s. (Sianesi, B., 2004. Rev. Econ. Stat. 86 (1), 133–155.)

The subject of the evaluation is a system with a wide array of different ongoing programmes which are held continuously through time and are open to all registered job-seekers. Sweden's active labour market policy is considered in its totality: all of the various programmes are aggregated into one 'programme' so that the treatment is any programme that a first-time unemployed person can join. The treatment is therefore starting a programme (in a given month). The effect estimated is the impact of joining a programme at a given time in unemployment compared to not joining it at least up to that time. This is different from the effect of participating in a programme rather than not participating, and from the effect of joining a programme at time *t1* rather than at time *t2.*

The data was obtained from two main sources, which reflect the programme component and the benefit component of the labour market policy. The first source is the unemployment register, which contains information on all unemployed individuals registered at the public employment office. It is available from 1991 onward and provides each individual's labour market status information over time (unemployed, on a given programme, temporarily employed, or similar), together with important personal characteristics of the job-seeker and of the occupation sought. The second source, available from 1994, is the unemployment insurance funds and provides additional information for those unemployed individuals who are entitled to unemployment benefits and assistance, in particular information on the amount and type of compensation paid out, previous wage and working hours.

The whole data set thus contains information about the duration of being in a labour market state, a set of demographic information (age, gender, citizenship), occupation being sought and human capital characteristics (specific and general education and occupation-specific experience), and, for entitled individuals, additional information on type of entitlement, unemployment benefits and previous working conditions. There is also information from an overall evaluation by the caseworker on the situation, character and needs of service of the job-seeker. This assessment relates to the job-seeker's degree of job readiness (judged to be able to take a job immediately, to be in need of guidance, or to be difficult to place) and to the job-seeker's preferences, inclinations, and sense of urgency (whether willing to move to another locality, looking for a part-time job or already having a part-time job).

Dynamic matching is applied in the following way: each individual enters the pool of those eligible for the treatment in the month that the individual registers as unemployed. Each month after the registration, individuals can be observed in three main states: still unemployed, participating in the programme or out of unemployment (including several options: finding a job, going back to full time education, inactive, etc..). The main comparison of interest is between individuals starting the programme and those who are unemployed until that time and do not participate in a programme yet.

Unemployment duration is set at a maximum of 18 months, since this captures 94% of all programme participants. A series of 18 regressions is estimated, corresponding to the maximum number of months after registration into unemployment when an individual can start a programme. Each regression models the probability of joining a programme in month *u*, conditional on the observed characteristics and on having been unemployed for *u* months. The propensity score is estimated through these regressions just as in the static case, and it is used to perform nearest neighbour matching, imposing common support.

The main outcome considered is labour market status up to 5 years after the first registration as unemployed. Labour market status is divided into two main categories: registered and deregistered from the employment office. Individuals registered can either be in open unemployment (receiving or not receiving unemployment benefits) or participating in a programme. Deregistered individuals can either be in employment or in other states (in education, inactive, or no longer in the system).

The paper answers two main questions:

1) How do unemployed individuals who join a programme perform *on average* compared to a situation where they would have searched further in open unemployment? This is found by calculating averages of the monthly effects to obtain a general idea of the effects.

2) What are the treatment effects *by month* of placement? This is found by looking separately at the time series of the various treatment effects for different subgroups of the treated that is based on the time they have spent in unemployment.

The main results of the analysis can be summarized as follow:

Average effects:

1) Treated individuals are more likely to be on programme participation and to be on unemployment benefits for 4 years after joining that programme. No effects are found for the probability of being in open unemployment.
2) After the lock-in effect, which is typical of individuals participating in programmes for the unemployed, over the 5 years since the programme started, the treated have a 6% higher ‘average’ probability of being in employment.
3) A detailed analysis of the mechanism carried out to explain the apparent contradiction of points 1) and 2) shows that the positive effects on employment, mentioned in point 2), are due to the fact that the programme reduced the probability of being unemployed outside the official unemployment system. Consequently, participants experience higher employment rates (point 2), but when they do become unemployed, they are significantly more likely to be entitled to benefits (point 1).
4) No effect is found on the probability of being in education while a small negative effect is found on the probability of being inactive.

Effects by month of placement

1) Looking at the effects by month of placement also helps to explain the apparent contradiction mentioned above.
2) Starting the programme within the first 6 months of being unemployed is found to have positive effects on employment, while starting the programme after 15 months of being unemployed is found to have no effect on employment.
3) On the other hand, starting the programme after 15 months has a larger effect on being compensated in the form of benefits, while starting earlier on has much smaller effects. This is visible evidence of the disincentive to work that is embedded in the institutional setup of the programmes: joining a programme greatly increases the probability of being in benefit-compensated unemployment through time, of participating in further programmes with the passing of time and more generally of remaining within the unemployment system. This is due to individuals who join the programme relatively late

(after 15 months from registering as unemployed) and who probably only joined the programme to continue to be eligible for unemployment benefits.

3.4. References

Sianesi, B., 2004. An evaluation of the Swedish system of active labour market programs in the 1990s. Rev. Econ. Stat. 86 (1), 133–155.

Sianesi, B., 2008. Differential effects of active labour market programmes for the unemployed. Labour Econ. 15 (3), 370–399

Fredriksson, P., Johansson, P. , 2008. Dynamic treatment assignment: the consequences for evaluations using observational data. J. Bus. Econ. Stat. 26 (4), 435–445.

Dynamic matching

PROS AND CONS

+ Dynamic matching can be used to estimate the impact of joining a programme at a given time compared to not joining it at that point in time. It is useful in the case of an ongoing programme. This means that it could also be used in settings where all individuals potentially receive the treatment at some point.

– It is data demanding: in addition to the requirements of the static matching (availability of characteristics determining programme participation), the evaluator should be able to observe monthly or quarterly information about the treated and non-treated populations, and the sample size should be large enough to perform reasonable matching in each of these time windows.

4. SYNTHETIC CONTROL METHOD

4.1. Background

The synthetic control method (SCM) is useful in evaluating the effects of policies or programmes that take place at aggregate level. For example, it can be used when a whole country is affected by a policy. The typical setting where this method is applied is when there is one treated unit and a few control units, and the aggregate outcome of interest is repeatedly observed through time, before the policy under investigation is implemented.

This method shares some similarities with difference-in-difference. While difference-in-difference is used when the number of observations (in the treated and control group) is large but the outcome is only observed at a few points in time, the SCM is used with aggregate values of the treated and of the control group when the number of observations is small but the outcome is observed many times.

This method was first developed by Abadie and Gardeazabal (2003) and many applications have since followed. In brief, the SCM is a statistical tool that has been increasingly used to evaluate the effect of an intervention on different *aggregate outcomes* (e.g. school dropout rate, educational attainment, economic growth, employment rate, average income) where a *single* unit is subjected to a particular policy intervention. This unit is compared to a *synthetic control unit,* which is created artificially to resemble the characteristics of the treated unit as much as possible, and it is constructed as a weighted average of different untreated units. Untreated units contribute with a larger or smaller weight (or do not contribute at all) in the construction of the synthetic control depending on certain observable characteristics which are relevant for the outcome of interest.

This method has recently been adopted to evaluate the effect of a policy change/introduction targeting aggregate entities such as geographic or administrative areas (countries, regions at NUTS2, or NUTS3 level). The opportunity of estimating policy impacts in cases where one unit (or a few) is treated is what makes SCM an important tool in the overall evaluation of interventions.

4.2. Method and data requirements

SCM is usually applied when there are *J+1* units (regions or countries) among which only one unit is exposed to the intervention of interest, and the remaining *J* units are therefore potential controls.

As in the usual counterfactual framework, the effect of the intervention is simply defined as the difference between the observed outcome of the treated unit after the policy implementation and what would have happened in the same unit had the policy not been implemented. As this latter outcome cannot be observed, one has to find something which closely resembles what would have happened if the policy had not been put in place. The method proposes building a fake (synthetic) control unit by combining all of the available potential control units, which are defined as ‘donor pool’.

In the absence of the treatment, the outcome for the treated is a function of observable (and non-observable) characteristics. The SCM is used to find a combination of weights of control units so that the weighted average of the characteristics of the control units closely resembles the characteristics observed for the treated unit in the pre-intervention period. This is done by minimising differences in the observable characteristics which are relevant to the outcomes. As a result, the pre-intervention outcome of the synthetic control also closely resembles the pre-intervention outcome of the treated unit.

The impact of the treatment is quantified by a simple difference of the treated unit against its synthetic cohort after the treatment. One of the great advantages of SCM is a very intuitive graphical analysis that shows the impact of a policy for the treatment and what would have happened to the same unit in the absence of the treatment (as represented by the synthetic cohort).

For example, a policy or a reform aimed at tackling early dropout from schools in a given country is to be evaluated. As the policy affects the whole country, *all* secondary schools are treated. Therefore to estimate the causal impact of the policy on the country's average dropout rate, a control group has to be found somewhere else *outside* the country. Simply comparing the dropout rate in the affected country to the same rates in other countries does not, of course, provide the causal effects of the policy: countries may be different in the outcomes for many other reasons. In addition, it is usually very difficult to find a single country with approximately the same characteristics of the affected country. The idea is that a combination of units should provide a better comparison for the treated unit than any individual unit on its own.

Figure 2 below provides a useful representation. The x-axis shows the timeline and y-axis the dropout rate in secondary schools of the country implementing the policy. The evolution of the dropout rate for the treated unit through time, i.e., country A, is displayed as a blue line while the red dashed line illustrates the counterfactual of country A. This latter is constructed as a weighted average of units from the donor pool, i.e., countries not affected by the policy. To form the closest match for the treated unit in the time period prior to treatment, similar potential control units are selected and assigned a positive weight. Observable characteristics considered to be those contributing to the dropout rate could for example be: expenditure in education as a function of GDP, presence of early tracking in the country, employment disadvantage for individuals with low educational achievement compared to individuals with a higher education, and lagged dropout rates. SCM builds a synthetic country, assigning different weights to the non-treated countries in a way that the four characteristics of the fake country are similar to those of the treated country. It is possible that many potential control units are not selected in the computation of the synthetic group due to non-trivial differences compared to the treated unit in the pre-treatment period. These units will be given a weight of zero in the construction of the synthetic control.

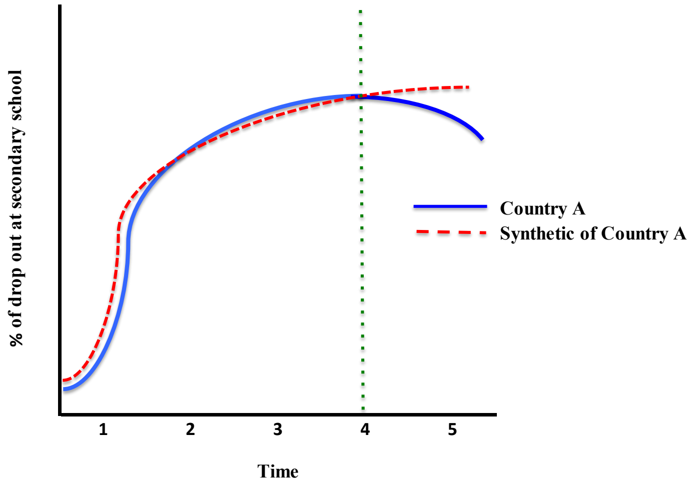

Figure 2: Synthetic cohort, graphical example

Figure 2 shows that the synthetic of country A closely resembles the evolution of the dropout rate in country A through time before the policy was implemented. Country A and its counterfactual follow a similar pattern at times 1, 2, and 3 (the pre-treatment period). Once the policy is introduced (at time 4), the treatment unit (country A) shows a sharp decrease in the percentage of pupils dropping out of school. The comparison between the two lines (red vs. blue) provides an initial hint of what would have happened without the change in the policy setup for country A.

Furthermore, the SCM allows a list of standard placebo tests to be carried out that help validate the intervention's causal effect. The placebos provide the probability of obtaining an estimate at least as large as that obtained for the unit representing the case of interest when the intervention is randomly reassigned in the data set (Abadie, Diamond, and Hainmueller, 2015). In this respect, it demonstrates what would have happened if the policy had been assigned at random to a given unit. The placebo effect should be close to zero. If, on the other side, the placebo policy shows a significant change in the outcome, larger than the effect estimated with the true model, this provides some indirect insights into the validity of the estimated model, meaning that the model is not valid and the estimated effect cannot be trusted.

Valid implementation of the SCM requires that the synthetic control closely matches the treated outcome during pre-treatment. If so, the comparison after the treatment provides a reliable estimation of the treatment's (policy) effect.

Additional requirements are:

Data must be available for several periods before the intervention of the policy in the treated unit and the pool of potential donor units.

1. Only the treated unit is affected by the policy intervention without any externalities to other units. In the example of the policy aimed at reducing dropout from secondary schools, this means that untreated countries are not affected by the treatment. The evaluator needs to check whether or not

countries used to build the synthetic control did not implement similar policies during the time period considered.

2. The policy intervention must not have any effect before it is enacted, which means that the outcomes (e.g. dropout rate in country A) do not vary due to the policy before the policy is implemented because of anticipation effects.

4.3. How it works in practice: An example from the academic literature

The impact of 'free choice': Family reforms in France and Belgium, a synthetic control analysis *by Federico Podestà (2017).*

This paper analyses the introduction of long-leave schemes and different actions to support childcare at home in France and Belgium between 1980 and 1990. This policy was characterised by women voluntarily participating in the programme. The author exploits the SCM to evaluate the impact of the reforms on the female labour participation rate (FLPR) at country level. His idea was to compare the treated countries (France and Belgium) with the synthetic units constructed using the following pool of countries: Australia, Austria, Canada, Denmark, Finland, Germany, Greece, Ireland, Japan, Luxembourg, Norway, Portugal, Spain, Sweden, the UK and the USA. The research question is to understand how the FLPR would have evolved had there not been these family programmes.

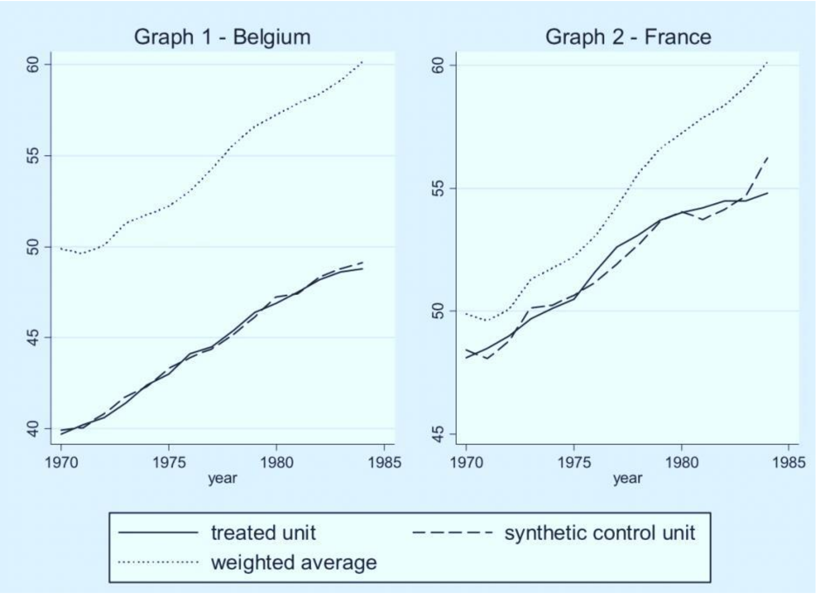

Figure 3 - Source Podestà (2017), page 8

The first step is to assign weights to each untreated country in order to build the two synthetic units that best mimic France and Belgium[6]. Using the routine available for many statistical software packages (STATA, R), the weights are optimised by selecting untreated countries similar to the treated ones and assigning a weight to them in the construction of the synthetic index. Indeed, the Belgium synthetic is constructed using a weighted average of Australia (11.4%), Canada (22.5%), Ireland (47.7%), Norway (7.1%) and the UK (11.4%), taking into consideration the following list of pre-intervention characteristics: percentage of workers in service, level of higher education attained by women, total fertility rate, number of weeks of maternity leave and unemployment rate. In addition, four lagged values of FLPR were included in the set of predictors (FLFP measured in 1984, 1978, 1974 and 1970). All of this data is available for the period 1970–2008, which makes the use of SCM appropriate. Similarly, this was carried out for France where it is important to underline the fact that the construction of the synthetic control unit for France was based on a different sample of donors (Ireland 15.9%, Japan 29.3%, Norway 8.3% and the UK 46.5%).

Figure 3 displays the trends in the Belgian and French FLPR compared to the synthetic unit and population-weighted average calculated on all units in the donor pool. Note that the distance between treated and population-weighted averages using all the donor countries is drastically larger than the distance between treated and synthetic controls, providing a graphical representation of the goodness of the counterfactual event. In other words, the synthetic units reasonably approximate the FLPR that would have been experienced by these countries through the pre-intervention period.

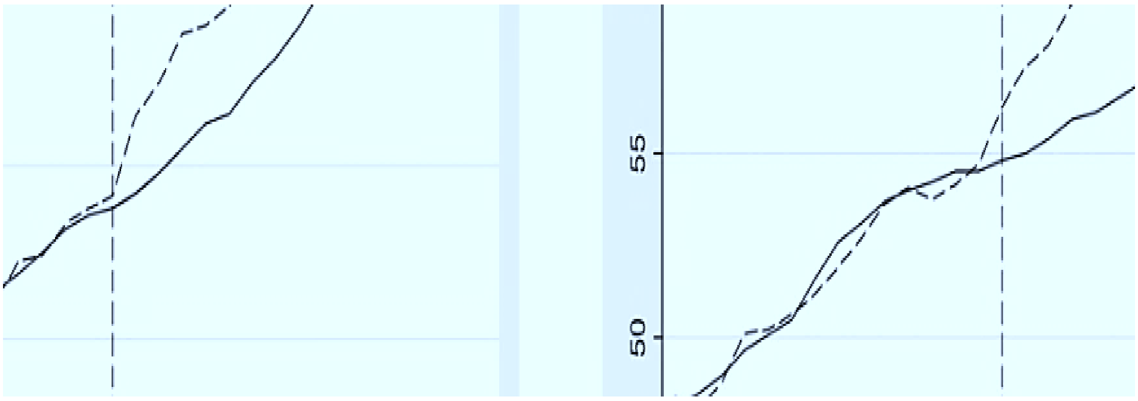

Figure 4 - Right panel Belgium; Left panel France; Source: Podestà (2017), 3 page 9

Figures 4 show that if both France and Belgium were not exposed to the reforms, their FLPR rates would have been higher than those actually observed. This is actually expected since the aim of the reform in both countries was to move both countries towards a more *familistic* model. It is worth noting that France (right figure) shows a slightly divergent path one year before the intervention (i.e. in 1984), which invalidates the identification of the causal path. However, the author finds that after also including Germany in the donor pool, the bias disappears due to the fact that Germany has a relevant weight that improves the fit.

[6] In this example there are two countries that are treated, but in the analysis they are considered to be two separate cases, and two separate estimates are computed.

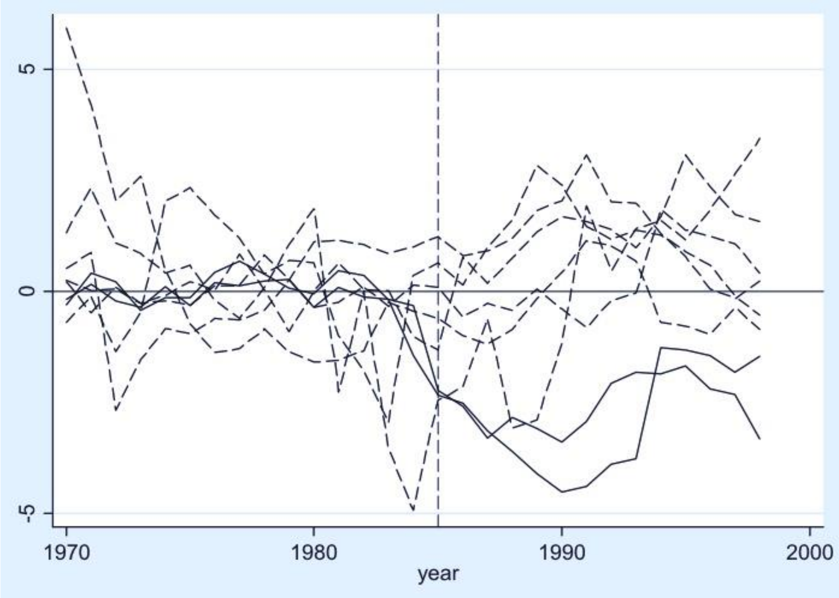

Figure 5 –Placebo exercise. Source: Podestà (2017), Figure 7 page 9

The authors also provide the placebo test. In the case of this paper, Figure 5 demonstrates that the gaps in the FLPR for Belgium and France (solid lines) are larger than the six placebos under a random year of reform implementation. The SCM exercise concludes that, if France and Belgium had not implemented this reform, women's participation in the labour market would actually have been higher.

PROS AND CONS

+ It can be used to evaluate policies affecting single units (one country, one region, …)

- To create a synthetic control, data must be available for a very long period before the intervention in the treated unit and the pool of potential donor units. The time series used in most academic papers is as long as 20 years.

4.4. References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. "Comparative politics and the synthetic control method." *American Journal of Political Science* 59.2 (2015): 495-510.

Abadie, Alberto, and Javier Gardeazabal. "The economic costs of conflict: A case study of the Basque Country." *American economic review* 93.1 (2003): 113-132.

Athey, Susan, and Guido W. Imbens. "The state of applied econometrics: Causality and policy evaluation." *Journal of Economic Perspectives* 31.2 (2017): 3-32.

Podestà, Federico. "The impact of 'free choice': Family reforms of France and Belgium, a synthetic control analysis." *International Journal of Social Welfare* 26.4 (2017): 340-352.

5. Conclusions

This document presents three methodologies that can be of help when carrying out a counterfactual impact evaluation.

The first two, sequence analysis and dynamic matching, rely on matching, where the evaluator seeks to find a control group of individuals as similar as possible to the treated group. Sequence analysis can be used in combination with standard matching techniques, when the data available goes back in time and allows recovering past history about relevant information on the population. Examples are past employment/unemployment history, when evaluating the effect of an active labour market programme; past hospitalisation history when evaluating the effect of a medical treatment, and so on. The past history is collected in the form of a sequence of events, and this is used to assess how similar a treated individual and a potential control individual are. While this can be of help, especially when the remaining observable characteristics available to the evaluators to perform the matching are not so informative or are very few, the amount of data required by this technique is substantial. One should be able to observe the status of each individual on a monthly/yearly basis for a reasonable amount of time before the intervention starts.

Dynamic matching can be used for ongoing programmes where at any time eligible individuals can start the treatment. In the academic literature, this method has been used especially to evaluate active labour programmes, which did not have a pre-set starting date and that were offered constantly to eligible unemployed individuals. With this method it is possible to estimate the effect of starting an intervention after *u* months of unemployment, rather than not starting it yet. Note that the comparison group includes individuals who will start the treatment later and those who will never start it. It is useful to assess if the effects of the treatment vary according to the starting month of the intervention. Similarly to static matching, dynamic matching requires that the evaluator is able to observe all the variables affecting selection into the treatment and outcomes, including also potential time varying variables. In order to be applied, a sufficient sample size in each of the months (or periods) of interest should be observed. If this is the case, then dynamic matching can really improve the estimation compared to static matching in the case of ongoing programmes, where potentially all individuals will receive the treatment at one point.

Finally, the synthetic control method was discussed. This method can be considered an extension of the difference-in-difference, where only one unit is treated. This can be the case of policies affecting a whole country or a single region. The method proposes to find a synthetic control unit, based on a weighted average of all the non-treated units (countries or regions). Then, by comparing the outcome of the treated unit to the one of the synthetic unit, it is possible to identify the causal effect of interest. To apply this method, the evaluator should be able to observe a long history of outcomes and control variables measured before the intervention. This is needed to identify the right set of control units to build the synthetic control. Additionally, it requires that the non-treated units are not affected by similar policies in the period after the intervention. Therefore, if similar policies are implemented in all the countries to be used as potential control, this method cannot be used to find a proper synthetic control unit.

Getting in touch with the EU

In person

All over the European Union there are hundreds of Europe Direct Information Centres. You can find the address of the centre nearest you at: http://europa.eu/contact

On the phone or by e-mail

Europe Direct is a service that answers your questions about the European Union. You can contact this service

– by freephone: 00 800 6 7 8 9 10 11 (certain operators may charge for these calls),

– at the following standard number: +32 22999696 or

– by electronic mail via: http://europa.eu/contact

Finding information about the EU

Online

Information about the European Union in all the official languages of the EU is available on the Europa website at: http://europa.eu

EU Publications

You can download or order free and priced EU publications from EU Bookshop at: http://bookshop.europa.eu. Multiple copies of free publications may be obtained by contacting Europe Direct or your local information centre (see http://europa.eu/contact)

EU law and related documents

For access to legal information from the EU, including all EU law since 1951 in all the official language versions, go to EUR-Lex at: http://eur-lex.europa.eu

Open data from the EU

The EU Open Data Portal (http://data.europa.eu/euodp/en/data) provides access to datasets from the EU. Data can be downloaded and reused for free, both for commercial and non-commercial purposes.

KE-04-19-627-EN-N

Publications Office
of the European Union