VTT Technical Research Centre of Finland

Selection of representative slices for generation expansion planning using regular decomposition

Helistö, Niina; Kiviluoma, Juha; Reittu, Hannu

Published in:
Energy

DOI:
10.1016/j.energy.2020.118585

Published: 15/11/2020

Document Version
Publisher's final version

Energy 211 (2020) 118585

Contents lists available at ScienceDirect

Energy

journal homepage: www.elsevier.com/locate/energy

Selection of representative slices for generation expansion planning using regular decomposition

Niina Helistö [a, *], Juha Kiviluoma [a], Hannu Reittu [b]

[a] Smart Energy and Built Environment, VTT Technical Research Centre of Finland Ltd, FI-02044, VTT, Espoo, Finland
[b] Data-Driven Solutions, VTT Technical Research Centre of Finland Ltd, FI-02044 VTT, Espoo, Finland

ARTICLE INFO

Article history:
Received 27 January 2020
Received in revised form
4 June 2020
Accepted 10 August 2020
Available online 22 August 2020

Keywords:
Clustering
Power system planning
Regular decomposition
Representative periods
Time series reduction
Variable renewable energy

ABSTRACT

In power and energy system planning tools, the temporal detail is often reduced by selecting representative slices out of longer time series. Various methods exist for the selection task, but they may prove slow or otherwise unfavourable in practical applications. Here, a generalized clustering algorithm, referred to as regular decomposition, is presented and applied to a power system planning study covering countries in the Northern Europe. The algorithm is compared with other selection methods, and the comparison is repeated with various number of representative slices and in three carbon price scenarios in order to provide more robust results. When selecting four weeks or more, regular decomposition is shown to perform relatively well compared to the other selection methods in terms of the total costs resulting from the power system model runs. When applied to inter-annual time series, regular decomposition is demonstrated to scale well. Although random sampling shows the most stable performance overall, the results indicate the need to test several methods for each system. Moreover, the results highlight the need to include net load peaks in the selected slices and to carefully estimate their position in the time series. A two-stage method for including net load peaks is presented.

1. Introduction

Planning of power and energy systems through optimization is a computationally demanding task and it has become more demanding with the increasing prevalence of variable power generation (VG) and different storage technologies. VG enforces the importance of correlated time series between multiple weather dependent variables including wind, photovoltaic (PV) and hydro generation as well as electricity and heat demand. Storage technologies pronounce the need for dynamic programming where temporal dependencies are fully considered. Meanwhile, uncertainty related to the energy resources, energy demands, technologies and technology costs further increases the computational challenge. Consequently, it is important to find ways to simplify the planning problem while maintaining sufficient accuracy. Aggregating similar consecutive time steps [1] and selecting representative slices, such as days or weeks, have been used to represent possible variations in the time series data using much less variables than full time series. Several methods for selecting the representative slices have been evaluated for this purpose.

The benefit of selecting representative slices compared to aggregating and averaging similar nonconsecutive time steps is that chronology is maintained between the time steps within each representative slice. Furthermore, representative slices can maintain the high temporal resolution and important variations in the original time series. The selection of representative slices has to simultaneously provide an adequate representation of the load duration curve while including potential correlations between the load and VG output, as argued by de Sisternes and Webster [2]. They selected a fixed number of weeks from one year of net load time series by applying a brute force method that calculates all possible combinations of weeks in order to select the sample with the smallest root mean square error. Computational issues limited the number of selected weeks to five even though only one area was analyzed. It is clear that the brute force method is not a good solution for multi-area systems if variations in each area are to be captured.

Abbreviations: CCGT, combined cycle power plant; CCS, carbon capture and storage; O&M, operational and maintenance; OCGT, gas turbine; PV, photovoltaic; RD, regular decomposition; RMSD, root mean square difference; SBM, stochastic block model; SRL, Szemerédi's Regularity Lemma; VG, variable power generation.

* Corresponding author.
E-mail address: niina.helisto@vtt.fi (N. Helistö).

https://doi.org/10.1016/j.energy.2020.118585

When the number of separate time series profiles increases, a larger fraction of the full year has to be considered in order to eliminate biases caused by the chosen slices [3]. Even if model solutions are close to each other in terms of total system costs, there may be considerable differences in the technology portfolios, driven by extreme values. The extreme events of all areas are not necessarily captured by a smaller number of hours used to represent the full time scope.

As opposed to brute force methods, there are clustering methods for partitioning time series into clusters and selecting prototypes for representing the clusters. One of the most commonly used clustering method is the *k*-means algorithm, where each object is reassigned to the cluster with the nearest mean in an iterative process. Other similar methods include *k*-medians, *k*-centers, and *k*-medoids. In the *k*-centers and *k*-medoids methods, cluster members are selected to represent the whole cluster and, consequently, they maintain the original correlations between the time series profiles. Meanwhile, *k*-means and *k*-medians create a new representative from the cluster members and consequently achieve a better fit with the original data, but at the same time create less variable profiles. These four clustering methods have been compared by Schütz et al. [4] for building energy systems while considering heat and electricity demand as well as solar irradiation. The results showed that the *k*-centers and *k*-medoids methods are better in the sense that they need less representative days to produce a negligible annual cost error when compared to a full year reference scenario.

In case studies of three alternative energy supply systems, Kotzur et al. [5] showed similar results where the *k*-medoids algorithm outperformed *k*-means. They also evaluated a hierarchical clustering method [6], which joined clusters in an iterative process one by one while minimizing the additional Euclidian distance from the full year set. The hierarchical method performed comparatively with the *k*-medoids algorithm and was computationally more efficient.

In addition to the aforementioned clustering methods, previous research has evaluated random sampling, optimization, and heuristic methods for reducing the temporal detail of power system planning studies. Poncelet et al. [7] compared methods based on heuristics, hierarchical clustering, random sampling, and optimization. They also considered ramps of original time series as separate time series. The optimization-based method showed the best results, but can become computationally heavy with increasing number of profiles. Its high computational effort was also stressed in relatively small-scale energy system studies in the Spanish [8] and Italian [9] context, where the optimization approach outperformed *k*-medoids, *k*-means, and a simple averaging method. Out of the other approaches in Ref. [7], random sampling outperformed the hierarchical method and is easy to implement and computationally light.

Palmintier et al. [10] used stratified sampling for a distribution network problem. They used repeated resampling in order to evaluate confidence intervals for the selected samples. This kind of approach would be valuable also for generation planning problems especially when using multiple years of profile data, but would benefit from a computationally efficient sampling method.

Storage technologies that have a longer cycle than the duration of the representative slice will be misrepresented by the above methods, as has been highlighted in the case of an island system [5] and in the Spanish context [11]. In order to consider storage technologies properly, Tejada-Arango et al. [11] used a system states approach with a transition matrix between the states and applied a *k*-means clustering method to obtain the states and the matrix [12]. However, increasing the number of the states as well as including additional time series profiles will quickly increase the size of the problem. The authors also presented a variation of the classical representative day approach, with cluster indices that enabled storage continuity and storage level checks at regular intervals. Representative days were selected using *k*-medoids.

In the work where different clustering methods were applied to three alternative energy supply systems, Kotzur et al. [5] concluded that the clustering method itself is not as important as the system under study. Similarly, based on a power system planning study of the Great Britain, Pfenninger [13] concluded that the most suitable approach to reduce time resolution depends on input data and model constraint setup, although heuristic approaches appeared more stable than statistical clustering. The same study also highlighted that better methods are needed to deal with inter-annual variability with high shares of VG.

Table 1 presents a summary of the methods applied in previous studies. The literature review indicates that there is no convergence on any particular method and that the best method may depend on the circumstances. At the same time, the representative slice selection challenge remains central to energy system planning. Hence, a methodological advance, even if it applies only to a sub-set of cases, can be of high importance. This paper makes a contribution by testing a novel selection method and compares it with a number of existing methods to establish its potential merit. In more detail, the contributions are as follows:

- The present research proposes an application of a generalized clustering method, referred to as regular decomposition (RD), to a power system planning study covering countries in the Northern Europe. The method is used to select representative weeks concurrently from load, wind, and PV time series.
- A two-stage approach is employed to include possible extreme situations in the set of selected weeks. Some extreme situations in the power system, such as the net load peak, are affected by the planning decisions, and their position in the time series is difficult to determine *a priori*.
- The performance of the RD method is demonstrated against other selection methods and the comparison is repeated with various number of selected weeks, in three CO_2 price scenarios, as well as in multi-zone and multi-year settings in order to provide more robust results of the performance of the algorithms. The quality of the selected algorithms is tested with the Backbone energy systems modelling framework [14], which is employed to optimize investments in and operations of thermal power generation, VG, and storage in a greenfield power system.

The hypothesis is that regular decomposition improves upon the existing methods that select representative slices for generation expansion planning. The method is explained in Section 2. Section 3 describes the case study including the models employed in the case study. Section 4 shows the results, followed by a discussion in

Table 1
Summary of the application of representative slice selection methods in the literature.

Method	Previously applied in
brute force	[2]
heuristics	[7,13]
hierarchical	[7,13]
k-centers	[4]
k-means	[4,5,9,13]
k-medians	[4]
k-medoids	[4,5,8,9,11]
optimization	[7–9]
random sampling	[7]
stratified sampling	[10]

Section 5 and conclusion in Section 6.

2. Methods

This section describes the variant of the regular decomposition method that is employed in this study as well as shortly presents the other selection methods in the comparison, the procedure for evaluating the quality of the selected slices and the approach to including extreme situations in the set of selected slices.

2.1. Regular decomposition of graphs and matrices

There is a vast number of clustering algorithms, the most popular like *k*-means are fast and easy to implement. We have developed a method that can be seen as generalization of clustering. It is inspired and justified by Szemerédi's Regularity Lemma (SRL). According to SRL, all large graphs have a structure, regular partition, that is very useful in understanding properties of large graphs. Regular partition has a bounded number of parts and—unlike clusters—has non-trivial connectivity between those parts. The connectivity patterns are similar to uniformly random, thus revealing redundancy among members of parts.

The algorithm for finding regular partition, corresponding exactly to the SRL, is difficult to implement. However, a similar structure, which replaces regular pairs by random bipartite graphs, can be found by a short algorithm called regular decomposition (RD). RD can be extended to matrices and used in data analysis. Such ingredients like information theory and SRL can make RD a viable addition to usual clustering methodology in data analysis.

RD was originally developed in works [15–20] for generic matrices in various use cases. In particular, Ref. [15] studies scalability and tolerance to missing data; Ref. [16] originates the method and uses it to link prediction; Ref. [17] extends RD to weighted directed graphs; Ref. [18] suggests to use graph distance matrix for RD in large and sparse graphs or matrices; Ref. [19] extends the method to arbitrary matrices with non-negative entries; finally [20] justifies the use of Minimum Description Length Principle as a basis for RD. In papers [19,21], RD was used to aid modelling and segmenting multiple time series of electric power consumption in households.

The algorithm for RD can be seen as a new variant of stochastic block model (SBM), see extensive reviews of SBM field in Refs. [22,23]. The emphasis here is on a more rigorous foundation of SBM. Clustering and SBM could significantly benefit from links to SRL and information theory as is suggested in RD. SRL can indicate possible new applications of the results of RD, while information theory can be used to find the right number of components of RD partition or number of clusters.

Using information theory as a basis of RD, roughly speaking, means segmenting data into optimal number of classes in such a way that redundancy in data is maximally revealed. As a result, any member of a class is similar to the other members in the same class. Thus, RD can be used to find representative slices of time series, by selecting representatives from regular group.

The main object in the RD algorithm is an $n \times m$ data matrix D with non-negative entries. For simplicity, we consider a matrix D with integer elements. Non-integer matrices are treated similarly and the algorithm for finding RD is identical to the one described below, as demonstrated in a previous work [19]. The result is the same cost function in both cases.

In this study, a version of RD is employed that partitions rows of D into k regular groups, first described in Ref. [19]. The partition is described by an $n \times k$ binary matrix R. Each of its rows corresponds to a row in matrix D and has value 1 in the position that indicates in which part or cluster $\{1, 2, \cdots, k\}$ the corresponding row of D belongs. All other elements on that row are zero. For instance, if $k = 5$ and a row i belongs to group number 2, then the ith row of R is $(0, 1, 0, 0, 0)$.

Matrices R and D define the following $m \times k$ matrix, P, the rows of which are the column averages of each part of the partition R of the matrix D:

$$(P)_{j,\alpha} := p_{j,\alpha} = \frac{\left(D^T R\right)_{j,\alpha}}{n_\alpha}, n_\alpha = \sum_{i=1}^{n} (R)_{i,\alpha}, \tag{1}$$

where $1 \le j \le m, 1 \le \alpha \le k$, and super index T is matrix transpose.

The minimum description length (MDL) framework [24] is used to find the optimal partition R. The coding length of the matrix D is $-\log\mathbf{P}(D|M)$ rounded up to the nearest integer, where $\mathbf{P}(D|M)$ is the probability of drawing matrix D from a probabilistic model M. In RD, the probabilistic model M is the following. Rows of D are partitioned in k groups according to some R. If row i belongs to group α, the circumstance is denoted as $\alpha(i) = \alpha$. Each matrix element $d_{i,j}$ with $i : \alpha(i) = \alpha$ is thought as generated from a Poisson distribution, independently from all other matrix elements, with parameter (expectation) equal to $p_{j,\alpha(i)}$. As a result, with a fixed D, the probabilistic model is uniquely defined by the R-matrix, according to (1).

Use of Poisson distribution is not an assumption about the actual distribution of matrix elements. It is used just for modelling purposes, to have a kind of measure of the proximity of matrix elements (Kullback-Leibler divergence) in regular groups. This assumption should be compared with the SBM for binary matrices, where binomial distribution is used in a similar role.

Using the Poisson distribution, the minus log-likelihood is found for each matrix element:

$$\begin{aligned} -\log\mathbf{P}(d_{i,j}|R) &= -\log\left(e^{-p_{j,\alpha(i)}} p_{j,\alpha(i)}^{d_{i,j}} \Big/ n!\right) \\ &= p_{j,\alpha(i)} - d_{i,j}\log\left(p_{j,\alpha(i)}\right) + \log n!. \end{aligned}$$

The corresponding code length for the entire matrix is the sum of all such elements. After removing terms independent on R, the R-dependent part of the coding length $-\log\mathbf{P}(D|R)$ becomes:

$$L_k(D|R) = \sum_{1\le i\le n, 1\le j\le m} \left(p_{j,\alpha(i)} - d_{i,j}\log\left(p_{j,\alpha(i)}\right)\right),$$

which is called the cost function. The small difference of integer valued coding length and $L_k(D|R)$ was ignored, insignificant for large D.

Optimal partition corresponds to minimal coding length:

$$R^* = \underset{R}{\operatorname{argmin}} L_k(D|R). \tag{2}$$

In order to find the approximately optimal partition, a greedy expectation-maximization algorithm was employed, similar to one in Ref. [19]. It runs in stages $s = 0, 1, 2, \cdots$. This algorithm starts at $s = 0$, from a uniformly random partition R_0 of rows in k non-empty groups. The greedy algorithm is defined by the mapping $R_{s+1} = \Phi(R_s)$. Mapping Φ is iterated until a fixed point, R_s, $\Phi(R_s) = R_s$ is reached. For each row i at stage $s+1$ the new group is defined from:

$$\alpha(i)_{s+1} = \arg\min_{1\le\alpha\le k} \sum_{1\le j\le m} \left(p_{j,\alpha}(s) - d_{i,j}\log\left(p_{j,\alpha}(s)\right)\right),$$

which defines the mapping $\Phi(\cdot)$. This program can be written in a matrix form by introducing an $n \times k$ matrix $C(s)$:

$$\begin{aligned}
&\alpha(i)_{s+1} = \underset{\alpha}{\operatorname{argmin}} C_{i,\alpha}(s) \\
&C_{i,\alpha}(s) := \sum_{1\le j\le m} p_{j,\alpha}(s) - (D \bullet LogP(s))_{i,\alpha} \\
&(LogP(s))_{j,\alpha} := \log p_{j,\alpha}(s), \\
&1 \le i \le n, 1 \le j \le m, 1 \le \alpha \le k
\end{aligned} \tag{3}$$

and where the matrix $P(s)$ is computed according to (1) using matrix R_s. Usually k is much smaller than n. Thus the main computational burden is in computing $P(s)$, which involves multiplication of D^T with R or multiplication of an $m\times n$ matrix with an $n \times k$ matrix. Usually the number of iterations needed is not high, typically just few rounds to reach a fixed point of Φ. In case of a very large matrix D, a sampling approach can be used, it is sufficient to have few samples from each row group in order to be able to compute the P-matrix. If the number of columns is also very large, then only a sparse sub-matrix of P is usually enough to find row groups. Similarly, the method is robust on missing data. The P-matrix can be estimated despite some missing values of D and classification is done using existing values, for more details, see Ref. [15].

Obviously, the global optimum of the program (2) is a fixed point of $\Phi(\bullet)$. On the other hand, the greedy algorithm (3) finds a fixed point that does not necessarily correspond to the global optimum of (2). That is why the solution is to run the greedy algorithm several times, starting from different random partitions. The fixed point corresponding to the minimal cost function value is taken as an approximate global optimum.

The matrix form of the algorithm is convenient and easily transferable to various programming environments, allowing creation of compact codes with only few lines of code. A Python implementation of the algorithm has been made available [25].

2.2. Selection methods in the comparison

When selecting representative slices, candidate slices are first extracted from the time series. In this work, we consider 1-week candidates, i. e, the duration of each candidate slice is 168 h. Furthermore, a *search interval* is defined, which denotes the gap between the starting positions of the candidates (see Fig. 1). A search interval that is smaller than the duration of the candidates allows overlapping candidates to be evaluated.

We compare the following selection methods: full year as a base case, RD, a modified k-means algorithm, random sampling, and a brute force method. The methods are shortly described in the following:

- RD has been presented in Section 2.1. Two versions of it, with different search intervals, were used: one that evaluates candidates starting every 48 h and another that evaluates candidates starting every 168 h.
- The random sample method selects X samples, each with n randomly selected slices. It then calculates the root mean square difference (RMSD) to the full year duration curves for each sample and selects the sample with the lowest value. A value of $X = 1000$ was used for the number of samples and the search interval was 168 h. The calculation of the RMSD is explained in Section 2.3.
- The modified k-means algorithm is a variation of the k-means method used by Kotzur et al. [5]. It groups the candidates into n clusters, and a member closest to the mean of the cluster is selected to represent the cluster. A search interval of 168 h was used.
- The brute force method evaluates all possible combinations while minimizing the RMSD to the full year duration curves. Similarly to the RD method, two versions of the brute force method were employed: one with a 48-hr search interval and one with a 168-hr search interval.

The modified k-means algorithm and the RD method do not always converge to the same solution. Thus, the modified k-mean algorithm was run 100 times and the best result in terms of the duration curve RMSD was selected. Similarly, RD was run 1000 times in order to find the best selection.

Most of the algorithms (RD with a 168-hr search interval, the modified k-means algorithm, and random sampling) were used to select 2, 3, 4, 6, 9, and 12 slices from the time series, each with a duration of 168 h. In the brute force method, the calculation becomes exponentially larger as a function of the number of the slices that are to be selected, n. Therefore, the brute force method—as well as RD with a 48-hr search interval—was employed to select at maximum 4 or 6 slices from the time series, depending on the search interval.

In order to analyze the efficiency and scalability of the selection methods, they were also applied to select representative weeks from 35-year hourly net load time series.

2.3. Quality evaluation

The quality of the selected slices is evaluated by means of power systems models based on the Backbone energy systems modelling framework [14], using a unidirectional soft-linking approach [26]. First, a planning model was run to acquire a portfolio of power plants for each zone in a multi-zone test system. The planning model optimized the generation and storage investments based on the selected slices—or based on the full year time series profiles in the base case. Second, a scheduling model was run for a full year to see the variable cost consequences of each portfolio. The most important criteria for the evaluation is the cost difference to the

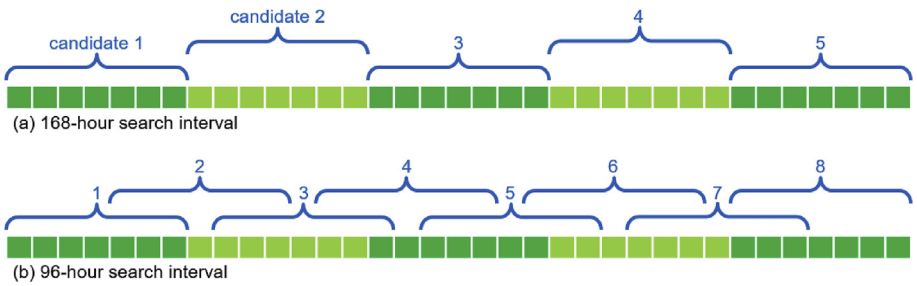

Fig. 1. Selecting 1-week slices from 5-week time series using different search intervals. Each green square represents a day. Search interval defines the gap between the starting positions of candidate slices. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

portfolio selected by the planning model based on the full year time series profiles. Another measure is to calculate the RMSD between the duration curves formed from the full year time series and the selected slices.

The RMSD between the duration curves can be calculated using different approaches. One approach is to treat zones separately and calculate the RMSD as follows:

1. Expand the selected slices to the same duration as the full year time series according to the weight of each slice.
2. For both the full year time series and the selected slices, normalize each time series category (load, wind, PV) separately, so that the average value across zones and time steps becomes 1:

$$p_{s,n,t} = o_{s,n,t} \times \frac{NT}{\sum_{n \in N} \sum_{t \in T} o_{s,n,t}},$$

$$\forall \{s, n, t\} \in SNT$$

where $o_{s,n,t}$ is the value in the original time series (category s, zone n, time step t), and $p_{s,n,t}$ is the corresponding value in the normalized time series. Due to the normalization, load, wind and PV will have the same weight in the end, while each zone is weighted by its relative contribution to the particular time series category.

3. Sort each time series in descending order.
4. Finally, calculate the RMSD:

$$\sqrt{\frac{\sum_{s \in S} \sum_{n \in N} \sum_{t \in T} \left(p_{s,n,t}^{\text{full}} - p_{s,n,t}^{\text{selected}} \right)^2}{SNT}}$$

In another approach, original time series are aggregated over all zones in each category before calculating the RMSD. Otherwise the approach follows the same procedure as presented above, with the exception that now $N = 1$. In this approach, each zone is still weighted by its relative contribution to the particular time series category, but the weighting is embedded in the aggregated time series. Each category is represented with an equally-weighted time series profile aggregated over zones.

2.4. Extreme situations

Simply selecting representative slices can lead to missing the extreme load and weather situations. Consequently, the planning model may not build enough capacity, and the scheduling model, which optimizes the operation of the system over the full time series, may end up in situations with energy not served. This has been observed in studies presenting a new subsampling approach [27], demonstrating the importance of storage and ramping dynamics [28], and introducing a framework for using clustering methods [29].

We manually included the week with the peak net load hour in each set of selected slices. The weight of this week was set to one. The selection methods were requested to select $n-1$ weeks to represent the remaining 51 weeks of the year. RD and the modified k-means algorithm were able to give different weights to the $n-1$ weeks, while the brute force method and the random sample method weighted all $n-1$ weeks with the same coefficient. In the end, each set of selected slices included n weeks with a total weight of 52.

The process for selecting the representative weeks was iterative in order to find the peak net load hour from the time series. In the first stage, VG capacity factor time series were multiplied by the average load of each zone and the time series were normalized such that the average value equaled one in each category. The zones were summed in each category and the resulting three time series were used as input to select five representative weeks using each of the methods. Next, the power plant capacities were roughly optimized with a planning model that used the five selected weeks as input. The VG capacity factor time series were multiplied by the installed capacities from these initial planning model runs in order to estimate the position of the peak net load hour. The corresponding week was removed from the full year time series that were given as input to the selection methods in the second stage. A similar two-stage approach was used by Hilbers et al. [27], who referred to it as importance subsampling. While they used the method with random sampling, we applied the iterative approach to all of the compared selection algorithms.

Using five representative weeks was assumed to be a reasonable compromise between accuracy and the number of planning model runs required in the first stage. However, in future applications it would be natural to use approximately the same number of representative slices in both stages.

3. Case study

The case study was designed to test the capabilities of the selection algorithms. To have a sufficient, yet manageable, challenge for all the algorithms, the test system has 12 zones each with correlated demand, wind power and PV time series based on the year 2011. The hourly time series data is from Northern Europe, including Germany, Poland, Estonia, Finland, Norway as three zones, Sweden as three zones, and Denmark as two zones. The time series are based on data provided by the transmission system operators, Nord Pool power market, and NASA MERRA reanalysis. Existing transmission connects the zones with each other.

The study starts from a greenfield system without existing power plants. In each of the 12 zones, there are five investment options, as shown in Table 2. Natural gas is the only fuel being considered and it can be burned in a combined cycle power plant (CCGT) or in a gas turbine (OCGT). The price of natural gas is assumed to be EUR 36 per MWh. Carbon capture and storage is possible in the combined cycle power plant with extra costs (CCGTCCS). It is assumed that 90% of the CO_2 can be captured and the cost of storage is EUR 10 per tonne of CO_2 stored.

In order to study the performance of the selection algorithms in low-renewable and high-renewable systems, three equally spaced levels for the CO_2 price were used: EUR 0 per tonne, EUR 50 per tonne and EUR 100 per tonne. The capacity margin assumption was 20% of the peak load, and it had to be fulfilled at every time step in the planning model. VG contributed to the available capacity according to its instantaneous generation. Instantaneous storage charging/discharging and power transfers on transmission links were also considered in the capacity margin constraint.

The quality of the selected algorithms was tested with the Backbone energy systems modelling framework, described in detail in Ref. [14]. The framework contains both a planning model and a scheduling model. The user can choose the level of detail in the representation of constraints in both Backbone models. Additionally, in the scheduling model, the level of detail can decrease towards the end of the rolling model horizon allowing more detail in the first hours where the final commitment and dispatch decisions are made. In this case study, the scheduling model determines the operations during one year using a rolling horizon with a 24-hr optimization period and an additional 8712-hr look-ahead period, where the temporal resolution decreases gradually from 1 h to 168 h. The planning model determines the number of units to build,

Table 2
Technology characteristics, adapted from Ref. [30,31].

Type	unit size	investment costs	annuity factor	fixed O&M costs	variable O&M costs	start cost	efficiency (min.)	efficiency (max.)	min. load	ramp up/down limit
	(MW)	(EUR/kW)	(/a)	(EUR/kW/a)	(EUR/MWh)	(EUR/MW)	(%)	(%)	(%)	(%/min)
OCGT	50	412	0.0858	7.423	4.50	43	40	45	20	10
CCGT	200	800	0.0858	26.000	4.00	43	58	63	40	4
CCGTCCS	178	1784	0.0858	44.720	7.80	43	51	56	40	4
wind	1	960	0.0806	11.340	1.22	0	100	100	0	–
PV	1	490	0.0750	7.810	0	0	100	100	0	–
battery	160[a]	135[a]	0.0806	1.620[a]	1.60	0	92	92	0	–

[a] For batteries, the unit of measurement for unit size, investment costs, and fixed O&M costs are MWh, EUR/kWh, and EUR/kWh/a, respectively. A battery with a capacity of 160 MWh is assumed to have a 80-MW charging capability and a 480-MW discharging capability.

but in this case study, the investment variables were continuously relaxed. Both the planning model and the scheduling model were formulated as linear programming (LP) models. Table 3 summarizes the implementation of different cost components and constraints in the model runs. Storage state as well as the online status of OCGT and CCGT (CCS) plants were forced to be equal at the beginning and at the end of each simulated week in the planning model.

Power and energy system investment decisions are preferably made based on multiple years of time series data, in which case having an efficient method for selecting representative slices becomes increasingly important. For this reason, another case study was designed based on 35 years of time series data from Finland.

4. Results

This section presents the results of the multi-zone and multi-year case studies. The multi-zone case study results are divided into results focusing on the selection of the representative slices and results from the power system model runs.

4.1. Selected representative slices

The initial model runs resulted in two estimates of the peak net load position. Based on most of the planning model runs with initial week selections, the estimated net load peak occurred in the 8th week. However, as a result of all planning model runs with the initial brute force selections and two out of three planning model runs based on the initial RD selections with a 48-hr search interval, the net load peak occurred in the 46th week of the year.

Figs. 2 and 3 show the computation times and the RMSD results of the selection methods in the second stage. The brute force method with a 48-hr search interval performs well in terms of the RMSD but proves impractically slow after a few representative weeks. Random sampling is the fastest of the methods and results mostly in the smallest RMSD, when excluding the brute force

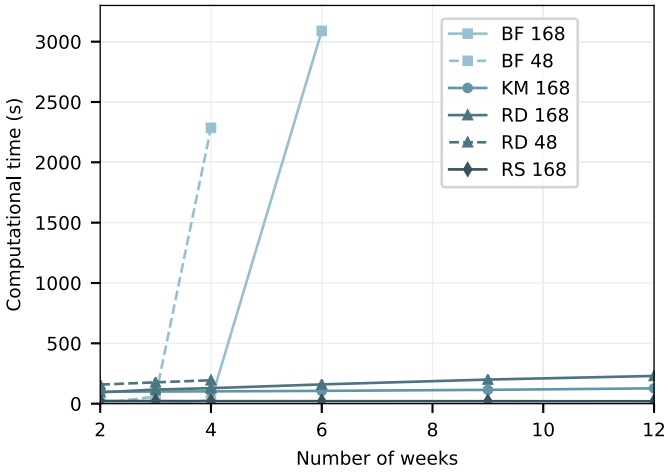

Fig. 2. Computation time when selecting representative weeks concurrently from three one-year time series of Northern Europe (load, wind, PV). The search interval was either 48 h or 168 h. The number of weeks does not include the estimated peak net load week. BF: brute force, KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

Table 3
Cost components and constraints in the model runs.

	Planning	Scheduling
Fuel use	no-load fuel use plus heat rate	no-load fuel use plus heat rate
O&M costs	EUR/MWh and EUR/MW/a[a]	EUR/MWh
Startup costs	EUR/MW	EUR/MW
Ramp costs	–	–
Investments costs	EUR/MW[b]	–
Penalties	loss of load, lack of reserve, lack of capacity	loss of load, lack of reserve
Energy balance	must be maintained, but has penalty variables	must be maintained, but has penalty variables
Reserve demand	single upward reserve	single upward reserve
Transmission	net transfer cap.	net transfer cap.
Online status	linearized	linearized
Min. operation	linearized	linearized
Start-ups	linearized	linearized
Shutdowns	linearized	linearized
Run-ups	–	–
Ramp constraints	yes	yes
Inertia	non-synchronous penetration limit	non-synchronous penetration limit
Capacity margin	net load plus margin	–
Forecasts	–	–

[a] EUR/MWh/a instead of EUR/MW/a for batteries.
[b] EUR/MWh instead of EUR/MW for batteries.

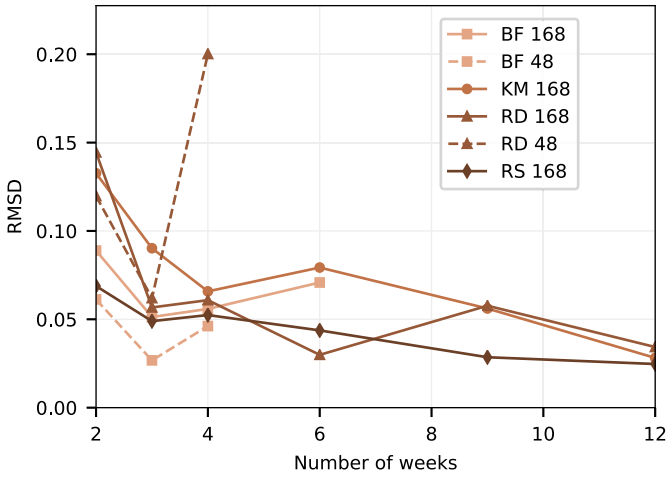

Fig. 3. Duration curve RMSD when selecting representative weeks concurrently from three one-year time series of Northern Europe (load, wind, PV). The search interval was either 48 h or 168 h. The number of weeks does not include the estimated peak net load week. BF: brute force, KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

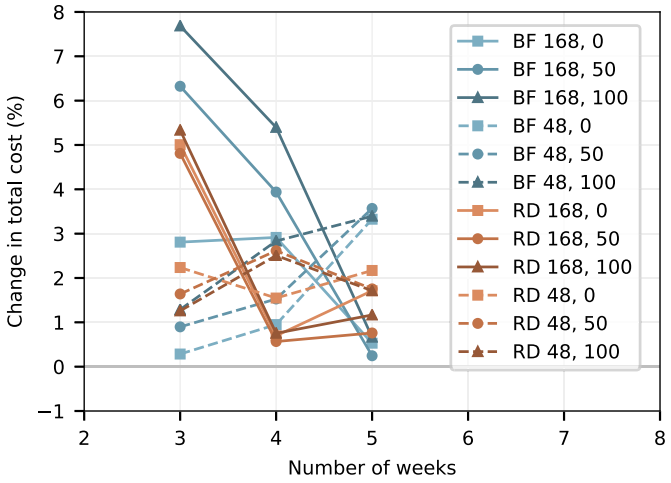

Fig. 5. Changes in total annual system costs compared to the full year simulations in the Northern European cases, with a search interval of either 48 h or 168 h. The legend states the selection method and CO_2 price (EUR/t), respectively. The number of weeks includes the estimated peak net load week. BF: brute force, RD: regular decomposition.

method. RD with a 168-hr search interval is slower than the modified *k*-means, especially at a relatively high number of selected weeks, but on the other hand, it has a comparable or smaller RMSD. Based on the limited number of tests shown in Figs. 2 and 3, the performance of RD declines when decreasing the search interval from 168 h to 48 h.

4.2. Energy system model results

Figs. 4 and 5 show the relative total costs resulting from the final multi-zone power system model runs. In most of the cases, the total cost results were 0–4% higher compared to the cost results of the full year runs, which served as base cases. The figures show many patterns similar to the RMSD results in Fig. 3, which suggests that the RMSD results are indicative of the quality of the selected slices. The relationship between the total costs and the RMSD results is further demonstrated in Fig. 6.

Fig. 4 shows that seven cases with a search interval of 168 h resulted in total costs that were more than 4% higher compared to the total costs in the corresponding base case. In these cases, the planning model was based on 3–4 weeks (incl. the estimated peak net load week). The number of weeks had to be increased to 5 in order to achieve total cost difference below 2% for all the methods.

In general, random sampling resulted in the lowest total costs (at maximum 2.2% higher than the corresponding full year case). One of the cases had even lower costs than the corresponding full year case. This may be due to certain modelling differences between the planning model and the scheduling model, such as having the capacity margin constraint only in the planning model. Random sampling also shows a small increase in cost results after 7 selected weeks, despite the decreasing RMSD (see Fig. 3). Nevertheless, the changes in the total cost results based on random sampling are minor after 7 selected weeks.

The total cost results based on the RD selections (with a 168-hr search interval) are comparable to the cost results based on the modified *k*-means algorithm and the random sampling method,

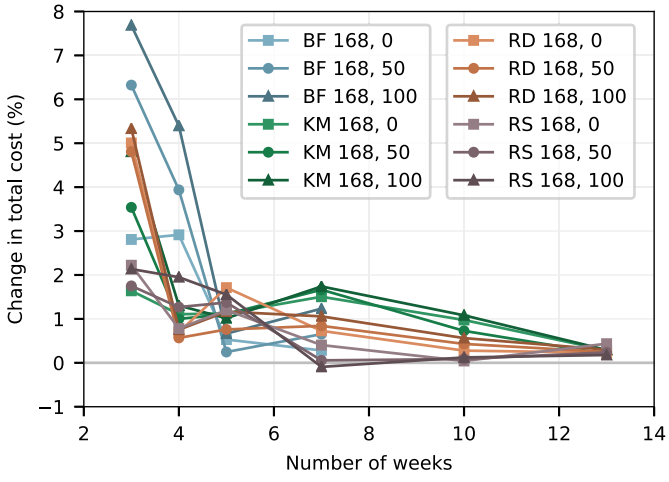

Fig. 4. Changes in total annual system costs compared to the full year simulations in the Northern European cases, with a search interval of 168 h. The legend states the selection method and CO_2 price (EUR/t), respectively. Total annual system costs in the full year simulations were EUR 61.24 billion (CO_2: EUR 0 per tonne), EUR 66.83 billion (CO_2: EUR 50 per tonne), and EUR 71.37 billion (CO_2: EUR 100 per tonne). The number of weeks includes the estimated peak net load week. BF: brute force, KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

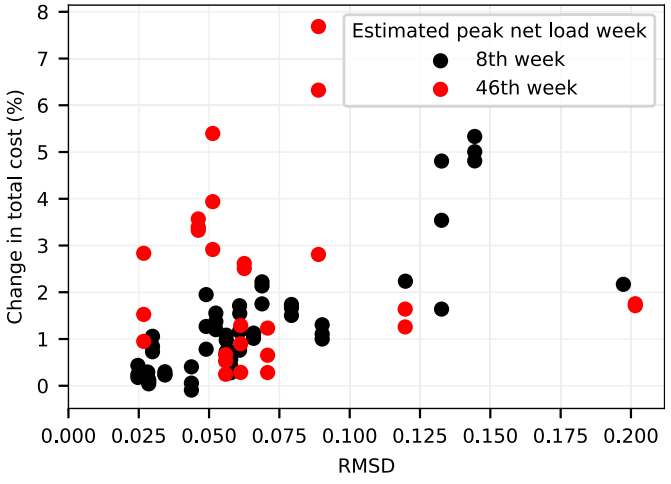

Fig. 6. Changes in total costs in relation to the RMSD in the Northern European cases.

when selecting 4–5 weeks for the planning model. When the number of weeks is increased to 7, the cost results based on the modified *k*-means are surprisingly high, but they start to decrease again when further increasing the number of selected weeks. At 13 weeks, the total cost differences between the cases based on these three selection methods are negligible.

Fig. 5 compares cases where the selection algorithms used different search intervals, but particularly it highlights that estimating the position of the net load peak requires careful attention. In most of the cases in Fig. 5, the estimated peak net load week was the 46th week of the year, and likewise, many of these cases resulted in a large amount of energy not served in the final scheduling model runs. However, there is also unaccountable behaviour in the costs of these cases, especially those based on the brute force method. It can be concluded that the results are likely to be very unstable if extreme situations are not correctly captured in the planning model.

These results could not demonstrate benefits from decreasing the search interval from 168 h when selecting representative weeks. Including only 3 weeks in the planning model (one of them being the estimated peak net load week) proved to be insufficient, as the total cost results varied considerably and were close to the optimal results most likely only by chance. Including 4–5 weeks in the planning model gives already more stable results, and after 7 selected weeks, all total cost results started to converge steadily. In general, the higher the CO_2 price was set—and the more the investments in VG grew—the larger the number of weeks had to be in order to achieve total costs close to the full year case. Optimal number of representative weeks or days will depend on the system at hand and the purpose of the study, as well as on the number of time steps that can reasonably be selected and included in the planning problem with the available computing power.

Although the total cost results were often quite close to each other, there were large differences in the cost component division as well as in the capacity and production results, as can be seen from Figs. 7–9, respectively. In particular, the investment results show large variation in the installed capacities of wind power, PV, and batteries (Fig. 8). The cases based on representative weeks resulted more often in larger amount of battery investments compared to the full year cases, whereas in the case of wind power and PV, full year results and mean results were relatively close to each other.

Furthermore, a comparison of the total cost differences (Figs. 4 and 5) and the division of investments (Fig. 8) demonstrates that there can be a plateau in the optimal total costs which can be achieved with various generation portfolios.

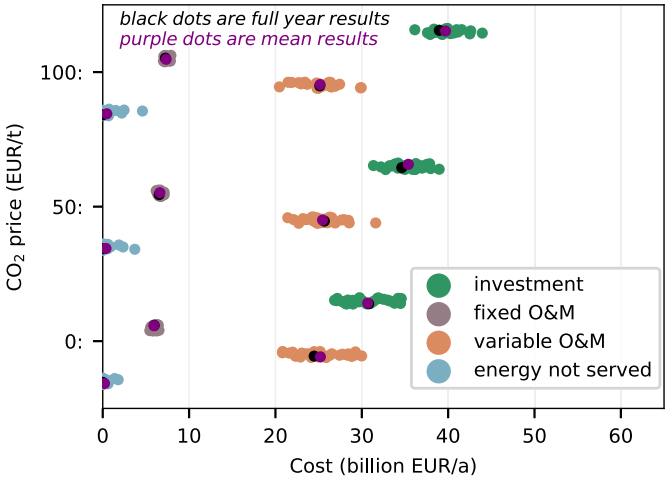

Fig. 7. Division of costs in the Northern European cases based on the planning and scheduling model runs.

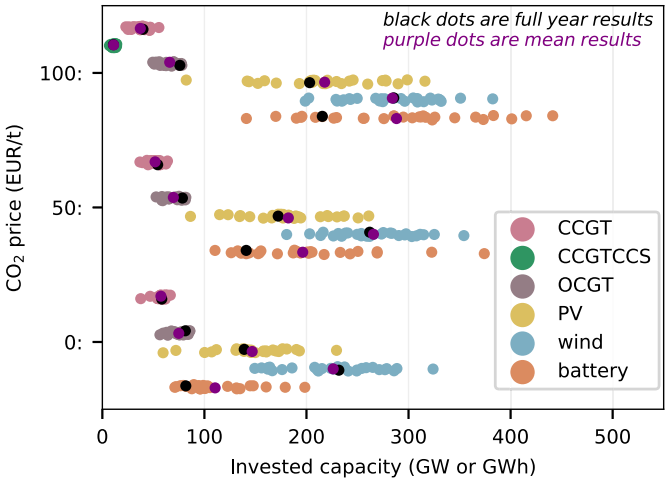

Fig. 8. Investments into new capacity in the Northern European cases based on the planning model runs. The unit of measurement is GWh for batteries and GW for other investment options.

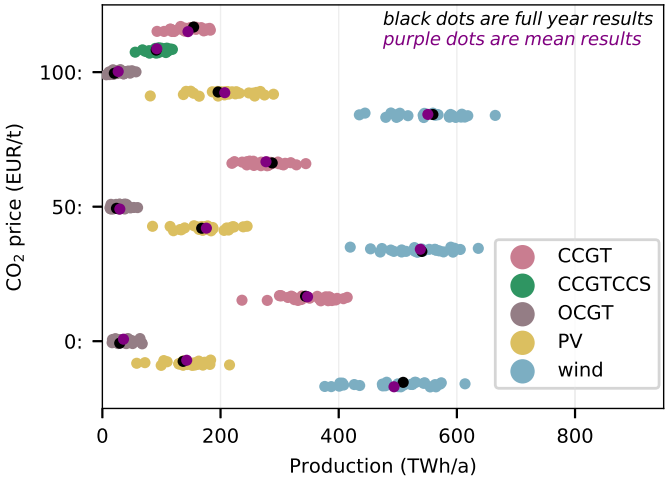

Fig. 9. Annual production per generation type in the Northern European cases based on the scheduling model runs.

4.3. Selecting representative slices from multi-year time series

The selection methods were also applied to select representative weeks from 35-year net load time series of Finland. Figs. 10 and 11 show that RD scales well compared to the modified *k*-means method, and it can be used for longer time series. Decreasing the search interval from 168 h to 48 h increases the computation time but does not show as significant improvements in the RMSD. The figure also shows that random sampling with $X = 1000$ samples is the fastest of the three methods and results in the smallest RMSD even when applied to multi-year data. However, although the analysis in Section 4.2 demonstrated that the RMSD between the duration curves is a good indication of the quality of the selected slices, it is not necessarily the best measure as it fails to take into account important patterns in the time series.

Therefore, Fig. 12 shows the total cost results based on the planning and scheduling model runs. The investment options were

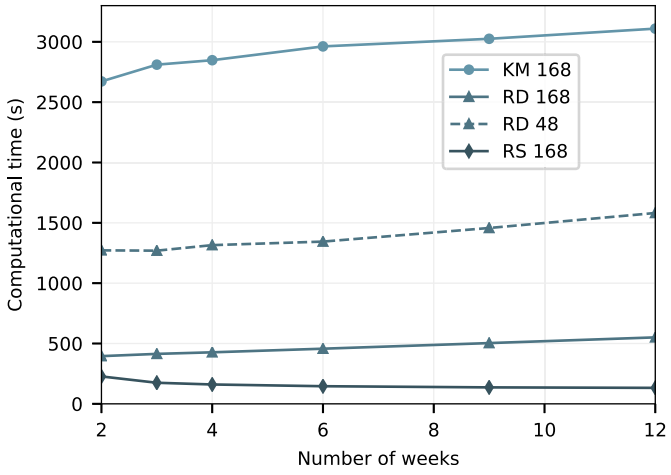

Fig. 10. Computation time when selecting representative weeks from the 35-year net load time series of Finland. The search interval was either 48 h or 168 h. The number of weeks does not include the estimated peak net load week. KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

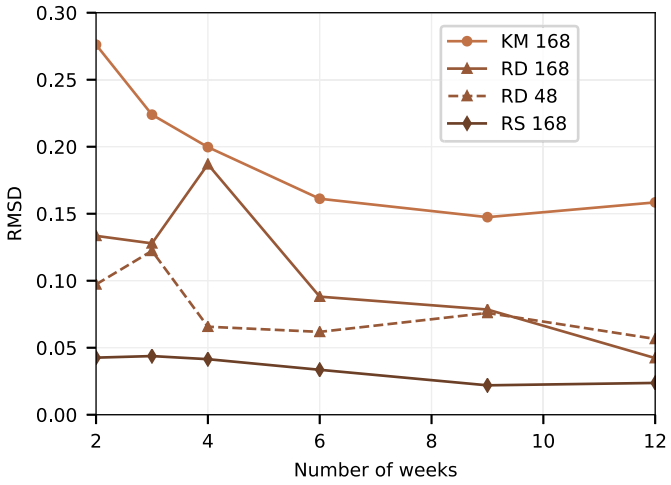

Fig. 11. Duration curve RMSD when selecting representative weeks from the 35-year net load time series of Finland. The search interval was either 48 h or 168 h. The number of weeks does not include the estimated peak net load week. KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

the same as in the Northern European cases, but now Finland was treated as an island system. In terms of total costs, RD with a 48-hr search interval and random sampling showed the best performance, as they resulted in relatively low and stable costs with all week selections. RD with a 168-hr search interval resulted in surprisingly high costs in the case of 7 selected weeks, and the costs resulting from the selections made by the modified *k*-means algorithm stayed on a high level still when the number of weeks was 10 or 13.

While RD scales well when it is applied to longer time series, the initial tests did not demonstrate that it would scale well when applied to multiple concurrent time series.

5. Discussion

In general, the RD method was shown to perform relatively well compared to the methods in the multi-zone test cases. However, RD was still outperformed by the random sampling method in terms of both the computation time and the total costs resulting from the

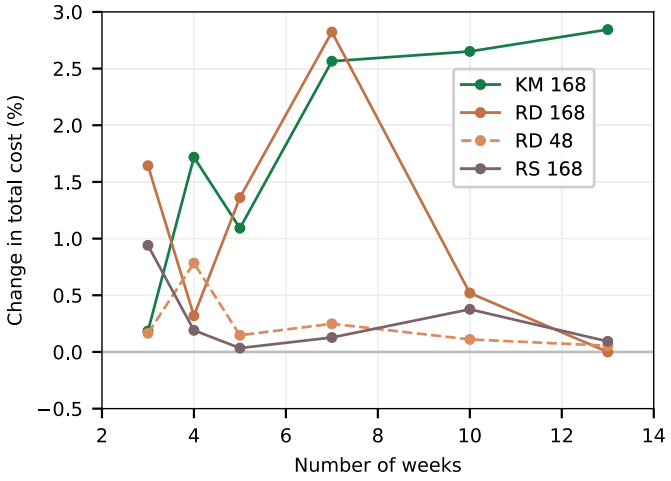

Fig. 12. Total annual system costs in the Finnish cases, relative to the case with the lowest costs (RD 168, 13 weeks: EUR 6.51 billion). The search interval was either 48 h or 168 h, and the price of CO_2 was EUR 50 per tonne. The number of weeks includes the estimated peak net load week. KM: modified *k*-means, RD: regular decomposition, RS: random sampling.

energy system model runs. We had expected that the random sample method would have difficulties in producing good results with larger data sets. However, it performed very well. One reason for this could be that cost effective power system investments, at least in the example power system, are economically sensitive to an adequate representation of the load and VG duration curves and less sensitive to ramps in the time series. The more advanced methods try to capture the variation as well, but, as a consequence, lose accuracy on the RMSD of the duration curves.

The results can be heavily affected by case and system dependence and the two-stage approach to estimate the peak net load period. In order to minimize the potential error resulting from case and system dependence, the comparison was repeated with various number of selected weeks, in three CO_2 price scenarios, as well as in multi-zone and multi-year settings. Using the two-stage approach resulted in misestimation of the net load peak position for some of the methods, but in another system or with more stages in the iterative approach, the results may have looked different. It should be noted that different modelling choices could have led to different relative performance of the methods and especially different RMSD and total cost difference values.

Moreover, the analysis showed that although there are large differences in the generation portfolios themselves, the total costs resulting from these portfolios can all be very close to the optimal total costs. Therefore, the selection methods should not be ranked solely based on the resulting generation portfolios. Depending on the purpose of the study, it may be important to consider various criteria, sometimes simultaneously, e.g. the total costs, total emissions, share of energy from renewable sources, and share of domestic energy production.

The RD method was demonstrated to scale well when it was applied to inter-annual time series, and the version that was able to evaluate overlapping candidates resulted in low and stable costs in all multi-year test cases. However, the method may not scale as well when applied to multiple concurrent time series. Further research is needed to ensure that the method is practical also in those applications. The sampling approach developed in Ref. [32] is anticipated to speed up computations for large matrices.

With increased sectoral integration, power system investment decisions cannot be made separately from other energy sectors [33]. Apart from batteries, many other storage options, such as

thermal storage tanks, seem more affordable and are relevant to consider [34]. This also leads to the need to better capture the behaviour of long-term storage in the models. Thus, an important direction for future research is to apply the methodology presented in this paper to multi-sector energy system studies, including links to heating and transport. When the method itself is efficient, applying it to a multi-sector energy system is straightforward.

6. Conclusion

A generalized clustering method has been presented, referred to as regular decomposition (RD), and its application in selecting representative weeks for a power system planning study has been demonstrated. The method was compared with other selection methods as well as using full year time series in the planning problem. In many real cases, using full year time series (or more) becomes computationally impossible, and practical methods are needed to reduce the time series data while still retaining important variations in them. Especially at high volumes of VG, more than 3 weeks should be selected and included in the planning model.

In general, regular decomposition was demonstrated to be an efficient method for selecting representative slices for generation expansion planning. However, the performance of the methods depends on input data and the number of slices to be selected. Although random sampling showed the most stable performance overall, the comparison demonstrates the benefits of testing multiple selection methods and number of slices. Regular decomposition scales well, and with further improvements, it could become a valuable method in the energy system planning toolkit.

The results also highlight the need to include net load peaks in the selected time slices with appropriate weighting and to not miss the final position of the net load peak that can be changed by the investment decisions. This may require, for example, an iterative process. This work employed a two-stage approach, but more stages could be introduced, where a new estimate of the net load peak would be included in the set of selected slices, until the scheduling model does not find situations with energy not served.

Credit author statement

Niina Helistö: Methodology, Software, Formal analysis, Investigation, Writing - original draft, Writing - review & editing, Visualization, Funding acquisition. Juha Kiviluoma: Conceptualization, Methodology, Software, Formal analysis, Writing - original draft, Writing - review & editing, Visualization, Supervision, Project administration, Funding acquisition. Hannu Reittu: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

The work was supported by the Academy of Finland project “Improving the value of variable and uncertain power generation in energy systems (VaGe)” (grant number 284973), which is part of the New Energy programme. N. Helistö has also received funding from the Jenny and Antti Wihuri Foundation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.energy.2020.118585.

References

[1] Pineda S, Morales JM. Chronological time-period clustering for optimal capacity expansion planning with storage. IEEE Trans Power Syst 2018;33(6):7162–70. https://doi.org/10.1109/TPWRS.2018.2842093.
[2] de Sisternes FJ, Webster MD. ptimal selection of sample weeks for approximating the net load in generation planning problems. jan 2013. ESD Working Paper Series No. ESD-WP-2013-03.
[3] Merrick JH. On representation of temporal variability in electricity capacity planning models. Energy Econ 2016;59:261–74. https://doi.org/10.1016/j.eneco.2016.08.001.
[4] Schütz T, Schraven MH, Harb H, Fuchs M, Müller D. Clustering algorithms for the selection of typical demand days for the optimal design of building energy systems. In: Proceedings of the 29th International Conference on Efficiency, Cost, Optimisation, Simulation and Environmental Impact of Energy Systems (ECOS 2016); 2016. p. 1–12. Portorož, Slovenia.
[5] Kotzur L, Markewitz P, Robinius M, Stolten D. Impact of different time series aggregation methods on optimal energy system design. Renew Energy 2018;117:474–87. https://doi.org/10.1016/j.renene.2017.10.017.
[6] Nahmmacher P, Schmid E, Hirth L, Knopf B. Carpe diem: a novel approach to select representative days for long-term power system modeling. Energy 2016;112:430–42. https://doi.org/10.1016/j.energy.2016.06.081.
[7] Poncelet K, Höschle H, Delarue E, Virag A, D'haeseleer W. Selecting representative days for capturing the implications of integrating intermittent renewables in generation expansion planning problems. IEEE Trans Power Syst 2017;32(3):1936–48. https://doi.org/10.1109/TPWRS.2016.2596803.
[8] Pinto ES, Serra LM, Lázaro A. Evaluation of methods to select representative days for the optimization of polygeneration systems. Renew Energy 2020;151:488–502. https://doi.org/10.1016/j.renene.2019.11.048.
[9] Zatti M, Gabba M, Freschini M, Rossi M, Gambarotta A, Morini M, et al. A novel clustering approach to select typical and extreme days for multi-energy systems design optimization. Energy 2019;181:1051–63. https://doi.org/10.1016/j.energy.2019.05.044.
[10] Palmintier B, Bugbee B, Gotseff P. Representative day selection using statistical bootstrapping for accelerating annual distribution simulations. 2017. In: IEEE Power and Energy Society Innovative Smart Grid Technologies Conference. Washington, DC, USA: ISGT 2017; 2017. p. 1–5. https://doi.org/10.1109/ISGT.2017.8086066.
[11] Tejada-Arango DA, Domeshek M, Wogrin S, Centeno E. Enhanced representative days and system states modeling for energy storage investment analysis. IEEE Trans Power Syst 2018;33(6):6534–44. https://doi.org/10.1109/TPWRS.2018.2819578.
[12] Wogrin S, Dueñas P, Delgadillo A, Reneses J. A new approach to model load levels in electric power systems with high renewable penetration. IEEE Trans Power Syst 2014;29(5):2210–8. https://doi.org/10.1109/TPWRS.2014.2300697.
[13] Pfenninger S. Dealing with multiple decades of hourly wind and PV time series in energy models: a comparison of methods to reduce time resolution and the planning implications of inter-annual variability. Appl Energy 2017;197:1–13. https://doi.org/10.1016/j.apenergy.2017.03.051.
[14] Helistö N, Kiviluoma J, Ikäheimo J, Rasku T, Rinne E, O'Dwyer C, et al. Backbone - an adaptable energy systems modelling framework. Energies 2019;12(17):3388. https://doi.org/10.3390/en12173388.
[15] Reittu H, Norros I, Bazsó F. Regular decomposition of large graphs and other structures: scalability and robustness towards missing data. In: Proc. Fourth International Workshop on High Performance Big Graph Data Management, Analysis, and Mining (BigGraphs 2017); 2017. Boston, USA.
[16] Nepusz T, Négyessy L, Tusnády G, Bazsó F. Reconstructing cortical networks: case of directed graphs with high level of reciprocity. In: Bollobás B, Kozma R, Miklós D, editors. Handbook of large-scale random networks. Berlin, Heidelberg: Springer Berlin Heidelberg; 2008. p. 325–68. https://doi.org/10.1007/978-3-540-69395-6_8.
[17] Pehkonen V, Reittu H. Szemerédi-type clustering of peer-to-peer streaming system. Proceedings of the 2011 International Workshop on Modeling, Analysis, and Control of Complex Networks. San Francisco, California: Cnet ('11); 2011. p. 23–30.
[18] Reittu H, Leskelä L, Räty T, Fiorucci M. Analysis of large sparse graphs using regular decomposition of graph distance matrices. December 10-13, 2018. In: IEEE International Conference on Big Data. Seattle, WA, USA: Big Data 2018; 2018. p. 3784–92. https://doi.org/10.1109/BigData.2018.8622118.
[19] Reittu H, Bazsó F, Weiss R. Regular decomposition of multivariate time series and other matrices. In: Fränti P, Brown G, Loog M, Escolano F, Pelillo M, editors. Structural, syntactic, and statistical pattern recognition (S+SSPR 2014), LNCS. vol. 8621. Berlin, Heidelberg: Springer Berlin Heidelberg; 2014. p. 424–33.
[20] Reittu H, Bazsó F, Norros I. Regular Decomposition: an information and graph theoretic approach to stochastic block models (jun 2017). URL, http://arxiv.org/abs/1704.07114.

[21] P. Kuusela, I. Norros, H. Reittu, K. Piira, Hierarchical multiplicative model for characterizing residential electricity consumption, Journal of Energy Engineering - ASCE 144(3).
[22] E. Abbe, Community detection and stochastic block models: recent developments (March 2017). doi:10.1561/0100000067.Emmanuel. URL http://arxiv.org/abs/1703.10146v1.
[23] Peixoto TP. Parsimonious module inference in large networks. Phys Rev Lett 2013;110:148701. https://doi.org/10.1103/PhysRevLett.110.148701.
[24] Grünwald PD. The minimum description length principle. MIT Press; 2007.
[25] Reittu H. Regular decomposition python code for simple graphs. URL https://github.com/hannureittu/Regular-decomposition.
[26] Helistö N, Kiviluoma J, Holttinen H, Lara JD, Hodge B-M. Including operational aspects in the planning of power systems with large amounts of variable generation: a review of modelling approaches. WIREs Energy Environment 2019;8(5):e341. https://doi.org/10.1002/wene.341.
[27] Hilbers AP, Brayshaw DJ, Gandy A. Importance subsampling: improving power system planning under climate-based uncertainty. Appl Energy 2019;251:113114. https://doi.org/10.1016/j.apenergy.2019.04.110.
[28] Scott IJ, Carvalho PM, Botterud A, Silva CA. Clustering representative days for power systems generation expansion planning: capturing the effects of variable renewables and energy storage. Appl Energy 2019;253:113603. https://doi.org/10.1016/j.apenergy.2019.113603.
[29] Teichgraeber H, Brandt AR. Clustering methods to find representative periods for the optimization of energy systems: an initial framework and comparison. Appl Energy 2019;239:1283–93. https://doi.org/10.1016/j.apenergy.2019.02.012.
[30] Energinet dk. Danish Energy Agency, Technology data for energy plants for electricity and district heating generation. jun 2019. Version number: 0003, Tech. rep., Copenhagen, Denmark, https://ens.dk/en/our-services/projections-and-models/technology-data/technology-data-generation-electricity-and.
[31] Energinet dk. Danish Energy Agency, Technology data for energy storage. Version number: 0002, Tech. rep., Copenhagen, Denmark (mar 2019), https://ens.dk/en/our-services/projections-and-models/technology-data/technology-data-energy-storage.
[32] Reittu H, Norros I, Räty T, Bolla M, Bazsó F. Regular decomposition of large graphs: foundation of a sampling approach to stochastic block model fitting. Data Science and Engineering 2019;4(1):44–60. https://doi.org/10.1007/s41019-019-0084-x.
[33] Hansen K, Breyer C, Lund H. Status and perspectives on 100% renewable energy systems. Energy 2019;175:471–80. https://doi.org/10.1016/j.energy.2019.03.092.
[34] Kiviluoma J, Rinne E, Helistö N. Comparison of flexibility options to improve the value of variable power generation. Int J Sustain Energy 2018;37(8):761–81. https://doi.org/10.1080/14786451.2017.1357554.