J Comb Optim (2011) 21: 306–329
DOI 10.1007/s10878-009-9249-2

The k-coloring fitness landscape

Hend Bouziri · Khaled Mellouli · El-Ghazali Talbi

Published online: 11 June 2009

Abstract This paper deals with the fitness landscape analysis of the k-coloring problem. We study several standard instances extracted from the second DIMACS benchmark. Statistical indicators are used to investigate both global and local structure of fitness landscapes. An approximative distance on the k-coloring space is proposed to perform these statistical measures. Local search operator trajectories on various landscapes are then studied using the time series analysis. Results are used to better understand the behavior of metaheuristics based on local search when dealing with the graph coloring problem.

Keywords k-coloring · Fitness landscape · Distance · Distribution of solutions · Time series

1 Introduction

Metaheuristics are widely used to solve various combinatorial optimization problems. The main reason for the interest in these techniques is their genericity. Indeed, a metaheuristic constitutes a general framework which can be applied to different optimization problems and thus present the ability to be adapted to a specific problem. Generally speaking, metaheuristics use move operators to "navigate" from a solution to another one in a search space and they stop if a satisfactory solution is found or if a

H. Bouziri (✉)
LARODEC-ISG, ESSEC, Tunis, Tunisia
e-mail: hend.bouziri@gnet.tn

K. Mellouli
LARODEC-ISG, IHEC, Carthage, Tunisia

E.-G. Talbi
LIFL, University of Lille 1, CNRS, INRIA, Lille, France

stopping criterion is satisfied. In addition, each metaheuristic framework provides its own technique to prevent the stagnation in local optima. Experiments performed on many instances of problems classified NP-hard, show the efficiency of these methods.

However, metaheuristics have the major disadvantage of a non guaranteed convergence. Their efficiency depends on many factors, such as the considered instance, the chosen cost function and the neighborhood operator. The research in the field of metaheuristics is focused on the adaptation of the framework to different combinatorial optimization problems without explaining why these search techniques perform so well. This insights us to stipulate that it is essential to study the search space structure, related to a given optimization problem, before developing a resolution method.

In this paper, the fitness landscape analysis is used in the study of the search space structure of the k-coloring problem. The fitness landscape is obtained by assigning to each point in the solution space a fitness value. Thus, the fitness landscape depends on the neighborhood operator, the fitness function and also on the considered instance. In this work, we perform the fitness landscape analysis for various k-coloring instances. In combinatorial optimization, the concept of fitness landscape has been used to understand the search process behavior of many problems such as in the resolution of the travelling salesman problem (Fonlupt et al. 1999; Boese 1995), the quadratic assignment problem (Merz and Freisleben 2000) or the knapsack problem (Travares et al. 2008).

This paper is organized as follows. In the next section, the fitness landscape analysis is discussed in more detail. Section 3 presents the parameters of the k-coloring fitness landscape and defines a distance on the k-coloring space. In Sect. 4, we present the experimental protocol. Results of the descriptive study of coloring distributions are discussed in Sect. 5. In Sect. 6, we show the use of the time series analysis in the study of k-coloring landscapes. Section 7 summarizes the measures used in our k-coloring fitness landscape analysis. We conclude in Sect. 8 and we propose some perspectives of this research.

2 Fitness landscape analysis

The concept of fitness landscape was originally introduced to understand evolutionary processes which are driven by specific operators (such as mutation and crossover). For the complete description of the model see Jones and Forrest (1995).

2.1 The landscape model

A fitness landscape L is defined as a 3-tuple:

$$L = (R, \phi, f), \tag{1}$$

where:

- R corresponds to the set of potential solutions that can be manipulated by the search procedure. The choice of the contents of this set constitutes the first step in solving a search problem. Solutions are then codified adequately.

- ϕ is an operator defined as $\phi : R \times R \to [0, 1]$. If $v, w \in R$, then $p = \phi(v, w)$ corresponds to the probability that v is transformed into w by a single application of the stochastic procedure represented by ϕ. In the evolutionary context, the neighborhood operator corresponds to classical genetic operators, i.e. crossover, selection and mutation.

  In our fitness landscape analysis, we define the neighborhood operator as a function $\phi : R \to R$, such that $\phi(v) = w$ means that the solution w is produced from v, by the application of the operator ϕ.
- f is a fitness function attaching to each genotype a value. In biology, the term *fitness* derives from the term "survival of the fittest" in the "natural selection".

To be able to deduce from the fitness landscape analysis, two properties have to be verified. First, the operator defining the neighborhood relation takes the same cardinality in input as it produces in output, so the resulting landscape is said to be *walkable* by the operator. Second, a landscape has to be *connected*, that is, there exists a path in the landscape graph between any pair of vertices.

2.2 Related works

The fitness landscape is a concept that is derived from the evolutionary theory. In the combinatorial optimization field, Fonlupt et al. (1999) work on the fitness landscape of the symmetric and Euclidean TSP. They show that the landscape is a massif central, i.e, a valley of local solutions with global solutions located at its bottom. Furthermore, the fitness distance correlation analysis performed on many instances, shows a large positive correlation between fitness and distance for the 2-opt operator if compared to the city swap operator.

Bachelet (1999) uses many statistical measures to study the fitness landscape of the QAP instances. This analysis reveals that most problems are unstructured: many instances have highly rugged landscapes and local minima are totally uncorrelated. Merz and Freisleben (2000) performed similar study by the use of other statistical measures to classify the QAP instances.

Many attempts have already been performed to analyze the search space of the graph coloring problem. In the following we present three different approaches in the search space analysis of the k-coloring problem.

2.2.1 A topological study

Hertz et al. (1994) proposed to study the performance of the tabu search algorithm in the resolution of the k-coloring problem. They start their analysis by generating the whole set of local optima solutions. Then, they provide statistical measures on the set of local solutions to illustrate the "topology" of the solution space.

This topological study on the whole set of local optima limits the size of problems that can be handled. In fact, it is difficult to enumerate all local optima for graphs with more than 20 vertices.

2.2.2 *Distance measures on the solution space*

Weinberg (2004) tried to give a classification of coloring instances according to two basic indicators: the entropy and the diameter. Furthermore, two distance measures between colorings are proposed, one is exact and the other is approximative. The second measure is used in the fitness landscape analysis basically to compute the diameter of the search space.

2.2.3 *Frozen sets*

Another related work on the search space analysis is that performed by Hamiez and Hao (2001) and Culberson (2000). They focused their study on the solution properties of the graph coloring problem. They discovered the existence of a particular set of vertices, that are always in the same color class, when solutions are generated. This set is called the *frozen set*.

The authors found that in practice, a coloring method can detect these frozen sets quickly. This implies that the resolution method *effort* is focused on the remaining pairs of nodes (that are not classified frozen).

Also, they enumerate the proportion of frozen sets for many instances, they found that it differs according to the considered instance. This should explain the fact that, for some graphs, solutions are multiple and different. While for others, solutions share common coloring information.

2.3 Measures on the fitness landscape

In our study, we propose some indicators to get an idea as complete as possible of the considered landscape. We distinguish between three types of measures:

- descriptive statistical measures that operate on solution distributions,
- the fitness distance indicator that attempts to study the potential relationship between fitness and distance to global solutions,
- correlation measures which aim at analyzing the ruggedness of landscapes.

To apply many of these measures we need a metric, related to the neighborhood operator, that provides the distance between any two solutions.

Let $d(s_i, s_j)$ be the distance between two solutions s_i and s_j in the search space. It corresponds to the number of neighborhood operator applications to obtain the solution s_j from the solution s_i. In topology, the distance is defined as follows.

Definition 1 (Distance) Let E be a set, we call *distance* on E, any mapping d: $E \times E \longrightarrow R+$, such that:

- $d(x, y) = 0$ if and only if $x = y$ (separative property).
- For all $x, y \in E$, $d(x, y) = d(y, x)$ (symmetrical property).
- For all $x, y, z \in E$, $d(x, y) + d(y, z) \geq d(x, z)$ (triangular property).

2.3.1 Descriptive statistical measures

In statistics, descriptive measures are commonly used to study distribution features such as the central tendency, the dispersion of individuals and the distribution shape. In what follows, we will adapt some of these measures to combinatorial optimization on search space.

Let P be a finite population of n solutions generated randomly. We associate to each solution s_i a fitness value denoted by f_i.

Normality assessment To show if a distribution is representative of the whole population, the idea is to measure the "shape" of the distribution relatively to the normal distribution. Accordingly, we use two measures: the skewness and the kurtosis statistics.

- The sample skewness is a measure of the lack of symmetry of a distribution. This statistic is given by the third central moment

$$\gamma_1 = \frac{\mu_3}{s_f^3}, \quad \text{with } \mu_3 = \frac{\sum_{i=1}^{n}(f_i - \bar{f})^3}{n-1}. \tag{2}$$

  Where, $s_f = \sqrt{\frac{\sum_{i=1}^{n}(f_i-\bar{f})^2}{n-1}}$ denotes the sample standard deviation. Since the normal distribution is symmetrical, it has a skewness value of 0. A positive skew value indicates rightward skewness, a negative skew value indicates leftward skewness.
- The sample kurtosis measure is given by the fourth central moment:

$$\gamma_2 = \frac{\mu_4}{s_f^4} - 3, \quad \text{where } \mu_4 = \frac{\sum_{i=1}^{n}(f_i - \bar{f})^4}{n-1}. \tag{3}$$

  A light-tailed distribution has fewer values in the tails than the normal distribution, and will have negative kurtosis. A heavy-tailed distribution has more values in the tails than the normal distribution, and will have positive kurtosis.

Diversity measures Measures on solution distributions can be classified into two categories: partition diversity measures and fitness diversity measures.

- *Partition diversity*: To measure the dispersion of solutions in the landscape, we use the mean of distances between all possible pairs of solutions $d(s_i, s_j)$ in the landscape. Given n the number of individuals in a population P, $\bar{d}$ is given by:

$$\bar{d} = \frac{2}{n(n-1)} \sum_{1} \sum_{j<i} d(s_i, s_j). \tag{4}$$

- *Fitness diversity*: The variability in point altitudes in the search space provides an approximative idea about the landscape "relief". To measure this diversity, we use the sample coefficient of variation cv which is given by the sample standard deviation s_f divided by the sample mean $\bar{f}$

$$cv = \frac{s_f}{\bar{f}} \quad \text{where } \bar{f} = \frac{\sum_{i=1}^{n} f_i}{n}. \tag{5}$$

2.3.2 Fitness distance correlation analysis

The fitness distance correlation (FDC) measures how much the fitness of a point correlates with its distance to a global optimum. The question is: *when will a search algorithm be effective at finding the global optimum*? Jones and Forrest (1995) suggest that the relationship between fitness (in an evolutionary context) and distance to a global optimum will have a strong effect on search difficulty.

To perform the fitness distance analysis, we need a sample of fitness values $F = \{f_1, f_2, \ldots, f_n\}$ and the corresponding set of distances to the optimal solutions $D = \{d_1, d_2, \ldots, d_n\}$. The correlation coefficient is given by:

$$r = \frac{\text{cov}(F, D)}{s_f s_d}, \tag{6}$$

where $\text{cov}(F, D) = \frac{1}{n}\sum_{i=1}^{n}(f_i - \bar{f})(d_i - \bar{d})$ and $s_d = \sqrt{\frac{\sum_{i=1}^{n}(d_i - \bar{d})^2}{n-1}}$.

The FDC analysis is founded on the following conjunctures:

1. Large positive correlations indicate that the problem is easy to be optimized since as the fitness decreases, the distance to the global optimum decreases also.
2. Large negative correlations indicate that the problem is "misleading" and the operator will guide the search away from the global optimum.
3. Near-zero correlations indicate that we should do a closer examination of the relation between fitness and distance through the use of a scatter plot of fitness versus distance.

One major limitation of the FDC analysis for real problems is that its computation requires the knowledge of the global optimum and the computation of the distance between the solutions, this cannot be done for a large variety of combinatorial optimization problems.

Another limitation of the fitness distance analysis is that all global optima have to be very close, otherwise, we can find different correlation results according to global solutions dispersions.

2.3.3 Correlation structure

The ruggedness of a landscape can be quantified by the correlation between adjacent configurations. The idea is to generate a random walk on the landscape via neighboring points. At each step, the fitness of the encountered solution is recorded. This way, a time series of fitness $f_1 \ldots f_n$ values is generated.

Autocorrelation function To measure the ruggedness, Hordijk (1995) proposed the *autocorrelation function* ρ_i, which is estimated by:

$$r_i = \frac{\sum_{t=1}^{T-i}(y_t - \bar{y})(y_{t+i} - \bar{y})}{\sum_{t=1}^{T}(y_t - \bar{y})^2}. \tag{7}$$

If r_i is close to one, then there is much correlation between two values i steps apart; if it is close to zero, then there is hardly any correlation.

Autocorrelation coefficient Angel and Zissimopoulos (1997) proposed another measure: *the ruggedness coefficient*. It is given by:

$$\xi = \frac{1}{1-\rho_1}. \tag{8}$$

This measure is based on the autocorrelation function of the nearest neighbor (ρ_1). The larger the correlation, the flatter the landscape is.

Box and Jenkins method (Box and Jenkins 1970) Once the time series of the "fitnesses" is obtained, a model can be built using the Box-Jenkins approach and thus we can make forecasts about future values, or simulate process as the one that generated the original data.

An important assumption should be considered here: the landscape has to be *statistically isotropic*, i.e., the time series of fitness forms a stationary random process. This means that the random walk is "representative" of the entire landscape, and thus the correlation structure of the time series can be regarded as the correlation structure of the whole landscape.

Hordijk (1995) used the time series analysis to determine the global structure of the fitness landscape of genetic operators. The purpose of this statistical approach is to find an ARMA model which adequately represents the data generating process. Hordijk found that the walk follows an AR(1) model and it is in the form:

$$y_t = c + \phi_1 y_{t-1} + a_t, \tag{9}$$

where a_t is a white noise which is a stationnary process such that the mean $E(a_t) = 0$, the variance $V(a_t) = \sigma^2$ and $\rho_k = \text{corr}(a_t, a_{t-k}) = 0$. This correlation structure implies that the fitness at a particular step in a random walk generated on this landscape, totally depends on the fitness one step earlier. Knowing the fitness, two steps earlier, does not give any extra information for the expected value of fitness at the current step. Furthermore, the value of the parameter ϕ_1 is the correlation coefficient between the fitness of two points one step apart in a random walk.

3 The k-coloring landscape parameters

Now, we will focus our study on the fitness landscape of the graph coloring problem. It is a key problem in combinatorial optimization, since it can modelize many theoretical and practical problems. Given a graph G with vertex set V and edge set E, a *proper* coloring of G is an assignment of colors to its vertices so that no two adjacent vertices in G have the same color. If the number of vertices k is fixed in advance, a proper k-coloring is a partition of V into k *independent* color classes.

The k-coloring problem can be seen as a decision problem, the question to be answered is whether for some positive integer k a proper k-coloring exists. It is well known that the k-coloring problem for general graphs is NP-complete and only for a few special cases polynomial time algorithms are known (Galinier 1999).

3.1 Representative space

It corresponds to the space of solutions and defines the set of points to be visited during the search process. Given a graph G with N nodes and k the number of colors, a k-coloring C corresponds to a set of k subsets representing color classes.

To avoid the *solution symmetry* in our analysis, we consider one solution as the set of colorings generating the same partition of vertices.

3.2 Fitness function

The goal of the search process, in our study, is to obtain a $proper\ k$-coloring, given a graph G and k colors to be assigned to each node. Our objective is to minimize the number of violated constraints, thus the cost function is given by:

$$f(C) = \sum_{i=1}^{k} |E(C_i)|, \tag{10}$$

where $E(C_i)$ is the set of edges of G having both endpoints in the same color class C_i. This function corrresponds to the value or *height* to be assigned to each point of the landscape.

3.3 Neighborhood operator

To move from a point to another one in the landscape, we need a neighborhood operator. In our investigation of the graph coloring landscape, we use the operator ϕ that changes the color of a node with another color. Thus, at each move a node is removed from a color class to be assigned to another one.

3.4 A distance on the k-coloring space

To perform the landscape analysis, we need an appropriate metric that measures the distance between solutions. In the definition of this distance, we have to take into account three basic considerations:

- This metric has to be related to the neighborhood operator since it computes the number of neighborhood operator applications, to obtain a solution, if we start with another one.
- The computation time of the distance has to be fast, since it will be computed for large instances.
- If we consider the hamming distance, each node is regarded individually with its color. Rather, we have to consider a node as a *member* of a color class. Thus, the hamming distance doesn't provide the adequate response of how far two k-colorings are. This idea is illustrated in Fig. 1, where the two colorings are similar, although, the hamming distance is equal to 4.

Consequently, we propose a distance based on partitions. The principle of this metric is to compute the hamming distance between color classes.

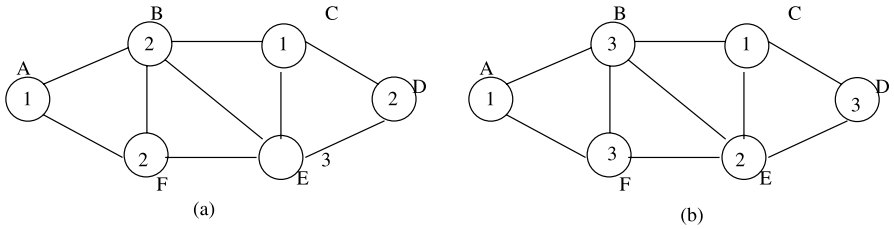

Fig. 1 The Hamming distance between the coloring (**a**) and (**b**) is equal to 4, although they are the same

3.4.1 Distance between color classes

A color class corresponds to the set of nodes of the graph having the same color. The purpose is to define a mapping that computes the distance between two node sets.

Definition 2 Given two sets A and B, we define a mapping d_s on the space of color sets, as follows:

$$d_s(A, B) = |A \cup B| - |A \cap B|. \tag{11}$$

That is, d_s computes the number of nodes that has *disappeared* or *appeared* when moving from A to B. Thus we have for instance:

- $d_s(\{1, 2\}, \{1\}) = 1$: the node 2 disappeared when moving from the first set to the second one (or appeared when moving from the second set to the first one).
- $d_s(\{1, 2, 3\}, \{1, 2, 4\}) = 2$: the node 3 disappeared and the node 4 appeared, when moving from the first set to the second one.

Proposition 1 *The mapping d_s is a* distance.

Proof It is obvious that d_s is symmetrical and separable. In addition, d_s verifies the triangular property, since we have:

$$|B| \geq |A \cap B| + |B \cap C| \quad \text{and} \quad |A| + |C| \geq |A \cup C|,$$

then

$$|A| + 2|B| + |C| \geq 2|A \cap B| + 2|B \cap C| + |A \cup C|.$$

So,

$$(|A| + |B| - |A \cap B|) + (|B| + |C| - |B \cap C|) \geq |A \cap B| + |B \cap C| + |A \cup C|$$

and

$$(|A \cup B| - |A \cap B|) + (|B \cup C| - |B \cap C|) \geq |A \cup C| \geq |A \cup C| - |A \cap C|.$$

Finally, $d_s(A, B) + d_s(B, C) \geq d_s(A, C)$.

It follows that the metric d_s is a *distance*. This distance can be used to measure the distance between two k-colorings. □

3.4.2 An approximative distance

Computing the distance between two k-colorings C_1 and C_2 can be seen as finding a minimum cost perfect matching in a weighted complete bipartite graph, where each node set corresponds to a k-coloring and each edge between two nodes V_i^1 and V_j^2 is weighted by the distance $d_s(V_i^1, V_j^2)$.

A *perfect matching* of a graph $G = (V, E)$ is a subset $M \subseteq E$ such that no two edges in M are adjacent and each node is incident to one edge in M. Since we use the same number of colors k, the graph that we use is bipartite and complete with $2k$ nodes and we can find a perfect matching by applying a hungarian method, with a complexity $O(k^3)$ (Kuhn 1956). If we take into account the calculation of distances d_s between color classes, we have a total complexity of $O(k^3 + N^2)$.

Despite the hungarian method is polynomial, it is still too time consuming for a fitness landscape study. In fact, in our descriptive analysis, we have to compute the distances between all pairs of solutions in several large distributions of large instances.

Hence, we propose a greedy method to measure the distance between two k-colorings. The method is given by the following algorithm which leads to an approximative distance d_a.

Algorithm Partition based distance.

data: $C_1 = (V_1^1, \ldots, V_k^1)$ and $C_2 = (V_1^2, \ldots, V_k^2)$.
result: $d_a(C_1, C_2)$.
begin
  $d_a(C_1, C_2) \longleftarrow 0$
  $cpt \longleftarrow 0$
  compute all the distances between each pair of color classes from C_1 and C_2
  repeat until $(cpt = k)$
    $d_s(V_i^1, V_j^2) \longleftarrow$ the smallest distance
    $d_a(C_1, C_2) \longleftarrow d_a(C_1, C_2) + d_s(V_i^1, V_j^2)$
    remove V_j^2 from C_2
    remove V_i^1 from C_1
    $cpt \longleftarrow cpt + 1$
  return $(d_a(C_1, C_2)/2)$
end

The algorithm starts by computing all possible distances d_s between color classes coming from C_1 and C_2. Then, at each iteration, the smallest distance is added to the total distance $d_a(C_1, C_2)$ and the color classes V_i^1 and V_j^2 are removed respectively from C_1 and C_2. The procedure stopped when the number of color classes is reached.

Consider the example of two 3-coloring of a graph with 13 vertices $\{a, \ldots, m\}$.

- $C_1 = (\{a\}, \{b, c\}, \{d, e, \ldots, m\})$
- $C_2 = (\{a, d, e, f\}, \{g\}, \{b, c, h, i, \ldots, m\})$

The bipartite graph of Fig. 2 gives the distances d_s as weights or costs of edges.

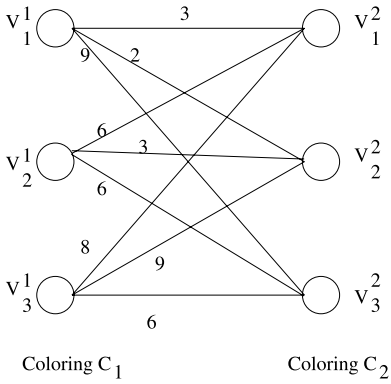

Fig. 2 The distance between two k-colorings corresponds to minimum cost perfect matching

The proposed algorithm will first match V_1^1 with V_2^2, then V_1^2 with V_2^1 and finally V_3^1 with V_3^2 for an approximative distance of 7, while the exact distance is 6 and is obtained by matching V_1^1 with V_1^2, then V_1^2 with V_2^2 and finally V_3^1 with V_3^2.

The complexity of the algorithm is $O(k^2 + N^2)$, where $O(N^2)$ is the time required to compute all distances d_s between nodes (color classes). The algorithm selects the minimal distance between k^2 possible d_s, then the corresponding pair of nodes is removed to restart the selection in the remaining $(k-1)^2$ distances, and so on. This leads to a complexity of $O(k^2)$.

In what follows, we use the proposed algorithm to compute the approximative distances between colorings. To detect the structure of the k-coloring landscape we use two basic measures. We first perform descriptive statistics on distributions of solutions. Next we use the time series analysis to study local search trajectories on the landscape.

4 Experimental protocol

Three families of graphs were chosen for our computational testing: Leighton graphs, flat graphs and random graphs. The instances are extracted from the second DIMACS Challenge.

Leighton instances are generated by a procedure which constructs graphs of known chromatic number. The flat graphs are proposed in Culberson (1996), they are generated in such a way as to reduce the variance of the degree of vertices (the number of adjacent nodes). As the flatness degree increases, the variation in degree may also increase; this variation should always be less than that of an equi-partite graph. The DSJC graphs are random, they are used in Johnson et al. (1993).

Columns 1 to 5, in Table 1, show for each studied graph, respectively, its name, the number of vertices, the number of edges, its chromatic number (or its best known lower bound when the chromatic number is unknown) (Desrosiers et al. 2004) and the best coloring found in the literature.

Table 1 The experimental protocol

Instances	Nodes	Edges	χ	Best k
dsjc125-1	125	1472	$\geq$5	5
dsjc125-5	125	7782	$\geq$10	17
dsjc250-9	125	13922	$\geq$30	44
dsjc250-1	250	64336	$\geq$8	8
dsjc250-5	250	31336	$\geq$11	28
dsjc250-9	250	55794	$\geq$35	72
dsjc500-1	500	24916	$\geq$12	12
dsjc500-5	500	125248	$\geq$16	49
dsjc500-9	500	224874	$\geq$42	127
flat300-28-0	300	21695	28	31
le450-5a	450	5714	5	5
le450-5b	450	5734	5	5
le450-5c	450	9803	5	5
le450-5d	450	9757	5	5
le450-15a	450	8168	15	15
le450-15b	450	8169	15	15
le450-15c	450	16680	15	15
le450-15d	450	16750	15	15
le450-25a	450	8260	25	25
le450-25b	450	8263	25	25
le450-25c	450	17343	25	25
le450-25d	450	17425	25	25

5 Descriptive statistical measures

Let P a population of n k-colorings generated randomly. We consider the corresponding fitness values and we propose statistical measures proposed in Sect. 3 to get, as possible, a complete idea about the landscape structure.

We start our descriptive investigation by the generation of an initial population of random colorings, for each instance. Then we apply a local descent on each individual to get a population of local optima.

Statistics are performed on 100 solutions generated randomly and on the 100 corresponding local optima. Statistical results are gathered in Table 2. Columns 1–6 show for each instance, respectively, its name, the number of colors used, the mean of distances, the coefficient of variation, the skewness statistic and the kurtosis statistic.

5.1 Normality assessment

For almost all instances, we note that the skewness value is near zero. This means that fitness distributions are symmetrical. Also, the results show that all the kurtosis statistics are practically equal to zero if they are negative and they are near zero if

Table 2 Descriptive statistical measures

Instance	k	$\bar{d}$		$c.v.$		Skewness		Kurtosis	
		Initial	Local	Initial	Local	Initial	Local	Initial	Local
le450-5a	5	344.667	344.512	0.0025	0.0027	−0.2129	−0.111	−0.319	−0.399
le450-5b	5	344.553	344.370	0.0023	0.0029	0.2330	−0.4345	0.0555	−0.1464
le450-5c	5	344.482	344.370	0.0028	0.0025	−0.054	0.255	0.0209	0.1635
le450-5d	5	344.3790	344.462	0.00168	0.00203	−0.0214	0.2596	−0.0517	−0.1617
le450-15a	15	386.969	387.359	0.0036	0.0037	−0.2927	−0.1053	0.02799	−0.5295
le450-15b	15	387.174	386.757	0.0035	0.0033	0.105	0.089	−0.3424	−0.5099
le450-15c	15	387.441	387.126	0.0032	0.0031	0.2083	0.0029	−0.6208	−0.3778
le450-15d	15	387.182	387.014	0.0023	0.0021	0.3370	0.2816	−0.6190	0.0013
le450-25a	25	383.153	382.928	0.0035	0.0032	0.3532	0.2067	0.4512	0.0748
le450-25b	25	383.381	383.132	0.0037	0.0031	−0.043	0.596	−0.3177	0.5133
le450-25c	25	382.411	382.689	0.0023	0.0023	0.2742	0.4774	−0.3177	0.5133
le450-25d	25	382.989	383.546	0.0037	0.0033	0.0293	0.3641	−0.2022	0.1064
flat300-28-0	28	240.607	240.583	0.0041	0.0039	0.2616	0.2455	−0.3436	0.0679
dsjc125-1	5	91.585	91.677	0.0046	0.0019	0.2217	0.1082	−0.2646	0.1512
dsjc125-5	17	93.005	93.210	0.0070	0.0066	0.8475	0.3762	−0.2628	1.7080
dsjc125-9	44	73.493	73.796	0.0055	0.0104	0.7372	0.3023	0.3270	0.0075
dsjc250-1	8	202.210	202.433	0.0042	0.0033	0.5458	0.5921	2.3746	−0.1138
dsjc250-5	28	197.361	197.877	0.0053	0.0045	−0.3952	1.7338	−0.3952	8.405
dsjc250-9	55	174.547	175.060	0.0044	0.0049	0.3582	0.3698	−0.2285	0.0684
dsjc500-1	12	429.810	429.879	0.0012	0.0010	−0.0304	0.0806	0.0983	−0.1340
dsjc500-5	55	410.235	404.839	0.0013	0.0014	0.2634	0.3380	0.2605	0.0382
dsjc500-9	128	335.927	335.655	0.0044	0.0030	0.7292	0.0350	0.6614	−0.5771

they are positive. This implies that the peakness of the distribution of fitness, relatively to almost all the instances, is practically similar to the peakness of the normal distribution.

Results of both kurtosis and skewness statistics, show, on the one hand, that the majority of fitness values are similar, which can be interpreted by the fact that altitudes are of the same level. On the other hand, we can stipulate that our investigation is adequate, statistically, since the normality conditions are assessed.

5.2 Stagnation phenomena

Results in Table 2 show that if the number of nodes increases, the diversity of solutions measured by the mean of distances $\bar{d}$ increases, regardless the number of color classes k. Also, we note that $\bar{d}$ is conserved when we generate the corresponding local optima. That is, if we generate an initial random population, we are rapidly trapped in local optima and thus the corresponding solutions aren't far from the initial ones. This can be interpreted by the presence of a multitude of non deep valleys in the landscape.

Furthermore, the coefficient of variation values are small near zero for almost all instances. This shows that the fitness values both in initial and local population are gathered near their means. This confirms the fact that locally altitudes have the same level. We can conclude then that locally the landscape structure of the graph coloring problem can be seen as a set of rugged plateaux.

This structure can explain the "stagnation phenomena" which is already detected in the resolution of the k-coloring problem, as in Galinier (1999) and in Galinier and Hertz (2004). They noticed that the objective function decreases dramatically in the early stages of the search and the search procedure generally uses most of its time trying to eliminate the last conflicting nodes.

Indeed, Fig. 3 shows that when coloring the le450-15c graph, the tabu search drops rapidly then stagnates. It requires four restarting points to reach the region where an optimal solution is detected. It is clear that if we omit the re-starting procedure, the search will stagnate for long time in the same region.

5.3 Frozen sets

Table 2 shows that if the graph size is maintained constant, the mean of distances remains basically the same for leightonian graphs. This can be interpreted as follows: even if the number of color classes changes, the same sets of nodes seem to be always together (frozen sets). Whereas, $\bar{d}$ changes for random graphs according to the instance. This can be explained by the absence of frozen sets, when random graphs are colored, or by their little proportion.

Hamiez and Hao (2001) and Culberson (2000), in their study on the solution properties of the graph coloring problem, showed the existence of a particular set of vertices that are always in the same color class when solutions are generated. This set is called the *frozen set*. In addition, Culberson (2000) concluded that these frozen same sets can be detected rapidly by greedy algorithms. This insights us to use random initial solutions when dealing with random graphs and to use a greedy method in the generation of initial solutions for leightonian graphs.

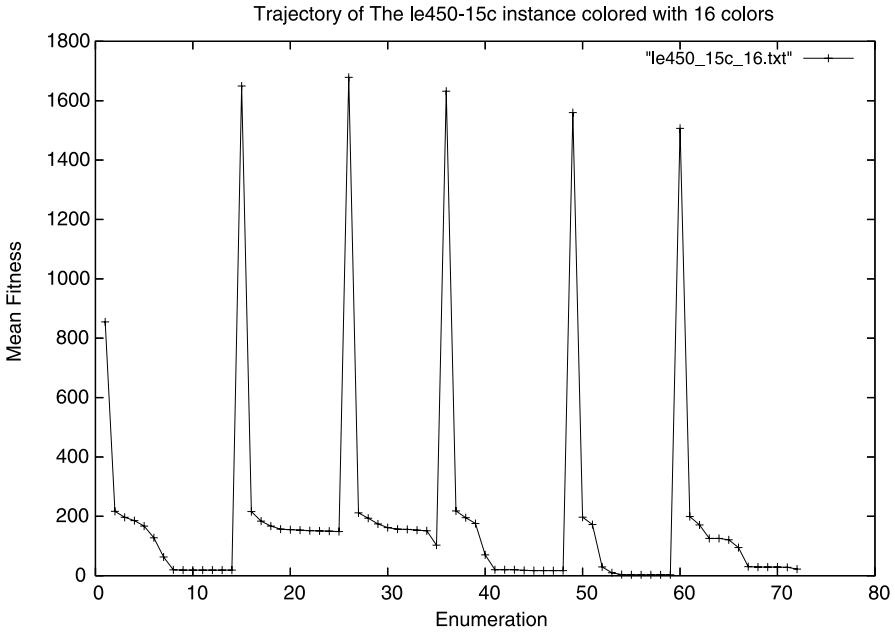

Fig. 3 The trajectory of the multi-starting tabu search algorithm in coloring le450-15c graph with 16 colors

To confirm our assumptions, we compare the behavior of the tabu search by the use of the greedy method DSATUR as initial solution generating method and its behavior when using random initial solution. We perform the tests on a leightonian graph (le450-5c) and a random graph (DSJC125-5).

Figure 4 clearly shows that the use of DSATUR as initial solution generation method improves the results of tabu search. However, for random graphs in Fig. 5, the use of DSATUR does not improve the effectiveness of the search in finding optimal solutions.

In this sense, many authors, as in Galinier and Hertz (2004), noticed that the use of a greedy method to generate initial solutions can be avoided since, in many cases, random initial solutions lead to similar results in time and qualities. But they do not explain this behavior; nor for which graphs it holds. They just concluded it by experiments.

6 Time series analysis

The aim of the use of the time series tool in our study is to modelize the walk or the trajectory performed by our neighborhood operator. This can give an idea about the landscape ruggedness. In the following, we begin by describing in details the Box and Jenkins procedure performed on the random instance of DSJC125-1 colored with five colors. Then we summarize results of the time series analysis on the set of instances used in the descriptive study of the previous section.

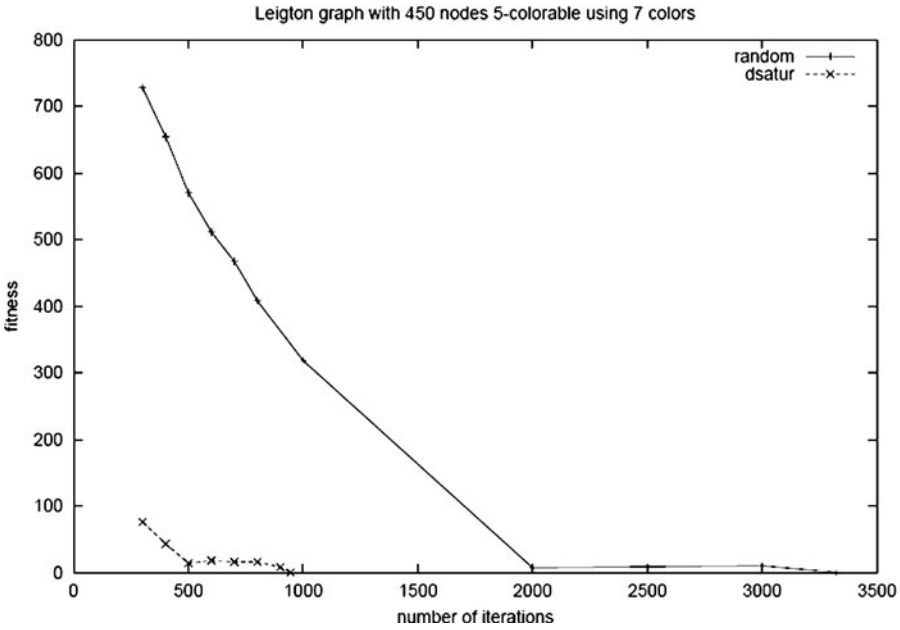

Fig. 4 The comparison of the search trajectory when using the DSATUR or random generation of initial solutions for the le450-5c graph

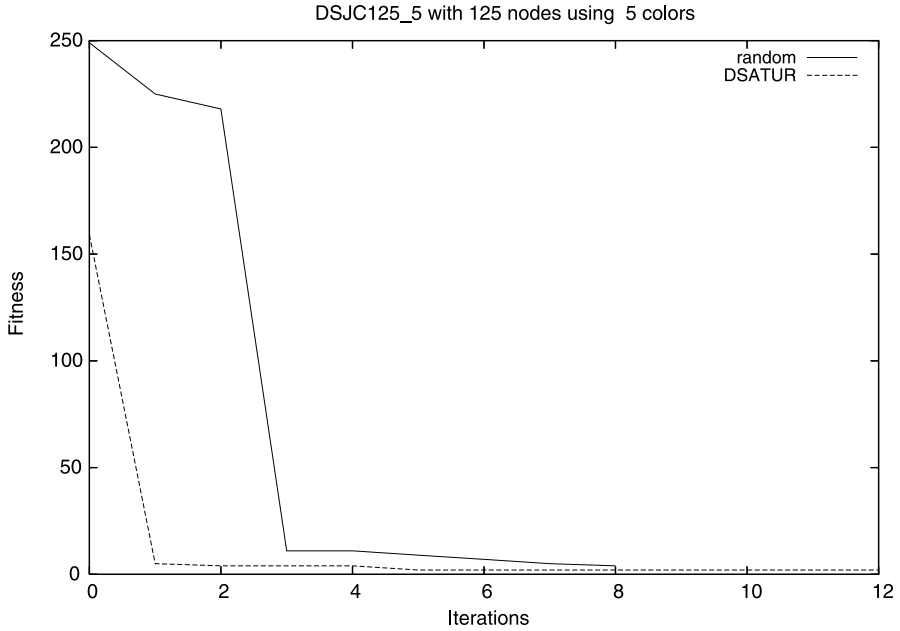

Fig. 5 The comparison of the search trajectory when using the DSATUR or random generation of initial solutions for the DSJC125-5 graph

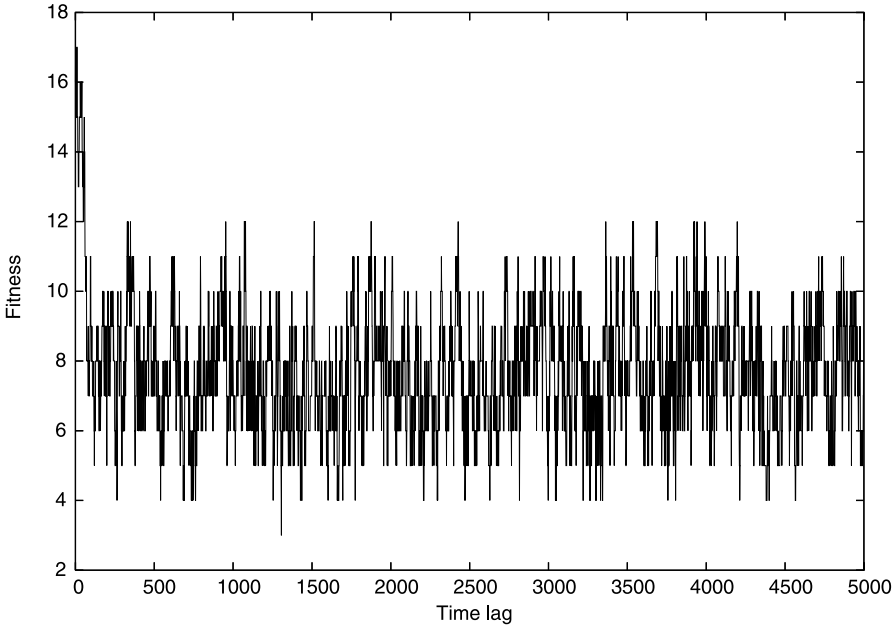

Fig. 6 The scatter plot of "observed" fitnesses

6.1 A detailed case study: the DSJC125-1 instance

This analysis starts by the generation of a random point (a random coloring of the graph) and by recording the corresponding fitness. The walk is then performed by the application of the neighborhood operator at each step of this walk. During this process, the fitness of the resulting points are computed and recorded. Once the time series of fitnesses is generated, we can apply the Box and Jenkins approach to provide a statistical model that represents the data-generating process.

6.1.1 Identification

This step aims at the determination of the model (or models) which can be used to represent the observed data. For this reason, we observe the scatter plot of Fig. 6. It shows that the series is stationary since observations are gathered around a constant mean value.

Furthermore, in Fig. 7, the correlogram (representation of the autoregressive functions or ACFs) is given together with the two-standard-error bound of $\frac{2}{\sqrt{T}}$ or ± 0.06 for the walk length $T = 1000$. The correlogram taper off to zero, so an AR(p) or an ARMA(p, q) should be most appropriate here. In this case we say that the ACFs die slowly because the values of the past carry over to affect the present.

Also, we compute the so-called Q or the Box-Ljung or portmanteau statistic which is based on the composite hypothesis that all the ACFs are equal to zero. Results on the DSJC125-1 instance show that all the Q-statistic are different from zero.

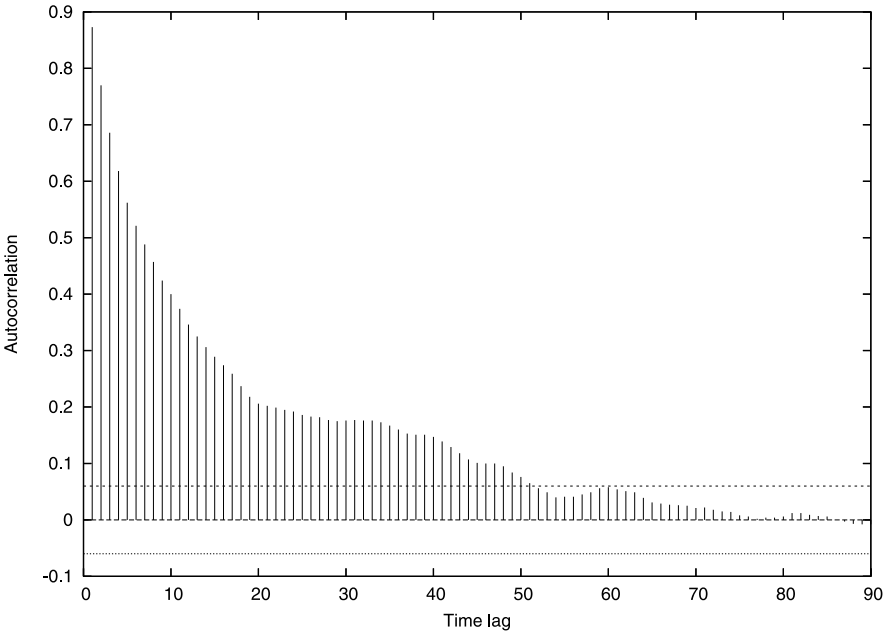

Fig. 7 The autocorrelation functions for the DSJC125-1 instance

To choose among the AR(p) and the ARMA(p, q) models, the partial correlogram, that presents the partial autoregressive functions or PACFs, is represented in Fig. 8. This plot shows that the first partial autocorrelation is equal to one and thus well outside the two standard error bound of ± 0.06. Whereas the other PACs are almost all within this bound.

Hence, the partial correlogram cuts off after one time lag and the AR(1) model is the appropriate choice to modelize the walk on the DSJC125-1 landscape.

6.1.2 Estimation

In general an AR(1) is given as:

$$f_t = c + \phi_1 f_{t-1} + a_t,$$

where the constant c is the mean of the time series. Estimation results are given by Table 3, where the constant $c = 7.688306$ and the correlation coefficient $\phi_1 = 0.872706$, so the model is as follows:

$$f_t = 7.688306 + 0.872706 f_{t-1} + a_t.$$

Also, Table 3 shows that all parameters are significant (t-*statistic* $\gg 2$). Furthermore the measure of "goodness of fit" R^2 is equal to 0.764931, so the estimated model is capable of explaining the observed data.

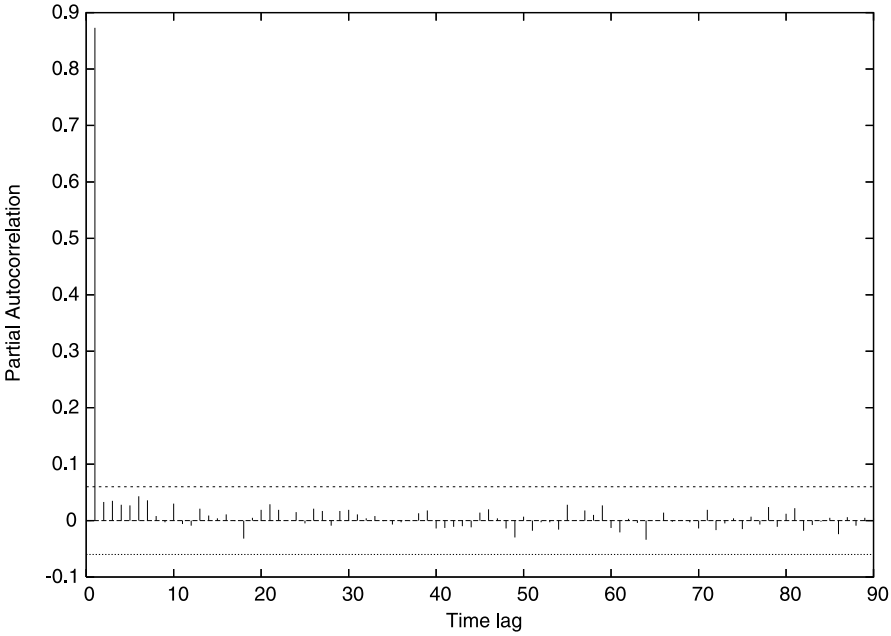

Fig. 8 The partial autocorrelation functions for the DSJC125-1 instance

Table 3 Estimation results for the DSJC125-1 instance walk

Variable	Coefficient	Std. error	*t-statistic*
C	7.688306	0.095821	80.23624
ϕ_1	0.872706	0.006845	127.5042

6.1.3 Diagnostic checking

To check if the estimated model is appropriate to be used, the model is fitted with the data and the autocorrelations of the residuals are computed. These residuals should be white noise, so all the autocorrelations should not be significantly different from zero.

The correlogram of the residuals in Fig. 9 shows that the first 200 autocorrelation functions of residuals are all within the two-standard-error bound. Thus, the first order autoregressive model has removed the significant autocorrelation in the data.

6.1.4 Model interpretation

This time series analysis performed on the DSJC125-1 instance is also applied to the other instances cited in Table 2. Results show that all the landscape instances can be modelized by an AR(1). That is, the correlation structure has the form:

$$f_t = c + \phi_1 f_{t-1} + a_t. \tag{12}$$

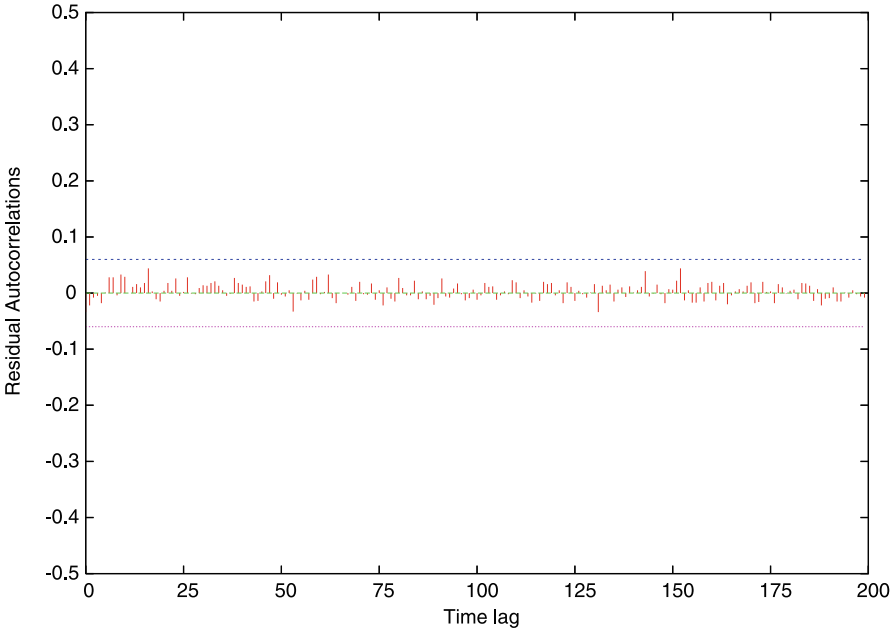

Fig. 9 The first 200 residual autocorrelations for the DSJC125-1 instance

The t-statistics of the all the estimated parameters are significant ($t\text{-}statistic > 2$). Furthermore, the value of R^2, a measure of goodness of fit of the model, indicates a high explanatory and predictive value of models.

The AR(1) model stipulates that the fitness, at a particular step, (f_t) in a random walk totally depends on the fitness one step earlier (f_{t-1}), and some stochastic variable. In these landscape models, the parameter (ϕ_1) is the correlation coefficient between the fitness of two points one step apart. The results show that it is high for all the instances.

6.2 An advanced study of correlograms

To give a more detailed analysis of the landscape correlation structure, we choose to observe correlograms in the same graphic to compare the correlation length of different landscapes.

Hordijk (1995) defines the correlation length as the largest time lag i for which the correlation between two points i steps apart is still statistically significant. That is, the quicker the correlogram drops to zero, the less the correlation length is.

First, we plot the correlograms by varying the size (we maintain the connectivity constant), we choose for this purpose the correlograms of three random instances with the same connectivity and various sizes: the DSJC125-1, the DSJC250-1 and the DSJC500-1 instances.

The corresponding correlograms in Fig. 10 all taper off to zero, but we see clearly that the correlation length increases as the size of the graph increases. What is ex-

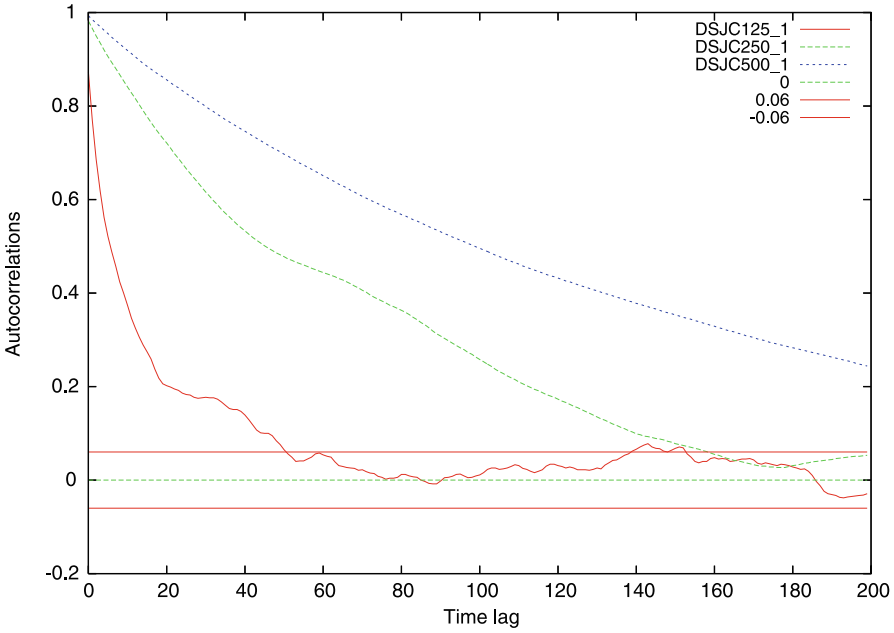

Fig. 10 Correlograms corresponding to random instances with different sizes

pected, since the larger the graph is, the more difficult it is to influence its fitness value in only one step.

Then, we vary the graph type. The correlograms of Fig. 11 show that the correlation length differs according to the graph type.

Results show clearly that the correlation length depends on the treated instance. This correlation length gives us an idea about the "flatness degree" of a given landscape. This can be useful in the determination of the length of the seach walk in local search methods. In Table 4, we provide results on the correlation length corresponding to the different tested instances, they are obtained by the use of S-PLUS statistical software.

7 Summary of landscape analysis

To get an idea as complete as possible about the k-coloring landscape, we have used many statistical indicators. These measures are gathered in Table 5.

Analysis results showed that for almost all the proposed instances, landscapes present a multitude of valleys and peaks and that local optima are distributed over the whole search space. Furthermore, altitudes of the peaks and depths of valleys seem to be regular. This should explain why local search techniques, that contain various strategies to escape from local solutions (valleys), are recommended. Also, diversity analysis showed the existence of frozen sets for leightonian graphs.

In addition, the time series analysis reveals that all the landscapes can be modelized by an AR(1), which means that in a current point, we can't see over the point

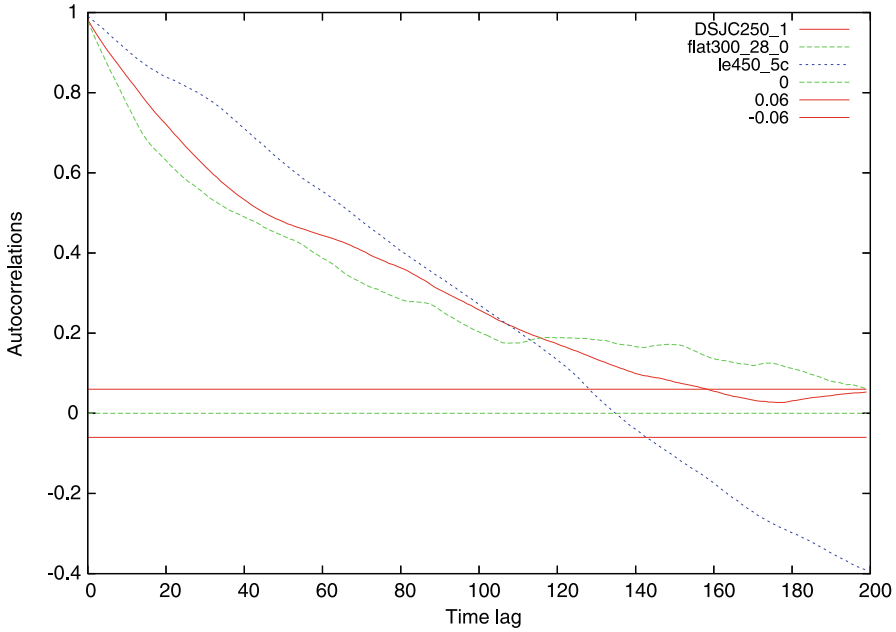

Fig. 11 Correlograms of different type of instances

Table 4 Correlation lengths corresponding to different k-coloring instances

Graph	k	Correlation length
Le450-5c	5	130
Le450-15c	15	214
Le450-25a	25	838
Le450-25b	25	396
Le450-25c	25	249
Flat300-28-0	28	202
DSJC125-1	5	65
DSJC125-5	17	151
DSJC125-9	44	370
DSJC250-5	28	678
DSJC250-9	55	150
DSJC500-1	12	452
DSJC500-5	55	847

following it immediately. The comparative study of the fitness time series, indicates basically that the correlation length depends on the specified instance, which implies different degree of flatness of the corresponding landscapes.

Table 5 Summary of measures used in the k-coloring landscape analysis

Statistical tools	Descriptive study			Time series analysis	
Data	k-coloring distribution			Search trajectory	
Measures	$\bar{d}$	$c.v.$	skew/kurt	Box-Jenkins	Correlograms
Landscape features	Partition diversity	Fitness diversity	Normality assessment	Correlation structure	Correlation comparison

8 Conclusion

The main objective of our research in this paper was the study of the fitness landscape of the graph coloring problem. For this purpose, we have proposed to use some descriptive tools to analyze several solution distributions. Also, we have used the Box and Jenkins approach aiming at modelizing trajectories generated by the adopted neighborhood operator. This is done for many instances of the second DIMACS challenge.

During our statistical investigation, we needed to define a distance between any two colorings C_1 and C_2, it computes the number of neighborhood operator applications, to obtain C_1 from C_2. Then, we have proposed a polynomial algorithm to approximate this distance.

As research perspectives, the results of this work can be fruitful in the elaboration of efficient search methods for the k-coloring problem. Also, we can perform a comparative study of different k-coloring landscapes by varying the neighborhood operator and/or the fitness function. This approach can also be applied to other grouping problems such as clustering and graph partitioning.

References

Angel E, Zissimopoulos V (1997) On the hardness of the quadratic assignment problem with metaheuristics. Technical Report, Laboratoire de Recherche en Informatique, University of Paris sud

Bachelet V (1999) Métaheuristiques parallèles hybrides: Application au problème d'affectation quadratique. PhD Thesis, Université des Sciences et Technologies de Lille, France, December 1999

Boese KD (1995) Cost versus distance in the travelling salesman problem. Technical Report UCLA computer science department, Los Angeles

Box GEP, Jenkins GM (1970) Time series analysis, forecasting and control, Holden Day

Culberson J (1996) On the futility of blind search. Technical Report 96-19, Department of Computing Science, University of Alberta, Edmonton, Alberta, Canada, July 1996

Culberson J (2000) Frozen development in graph coloring. Technical Report APES-19-2000, APES Research Group, February 2000

Desrosiers C, Galinier P, Hertz A (2004) Efficient Algorithms for Finding Critical Subgraphs. Les Cahiers de GERAD G-2004-31, April 2004

Fonlupt C, Robillard D, Preux P, Talbi EG (1999) Fitness landscape and performance of meta-heuristics. In: Voss S, Martello S, Osman I, Roucairol C (eds) Metaheuristics—advances and trends in local search paradigms for optimization. Kluwer Academic, Dordrecht, pp 255–266. Chapter 18

Galinier P (1999) Etude des métaheuristiques pour la résolution du problème de satisfaction de contraintes et de la coloration de graphes. Thèse de Doctorat de l'Université de Monpellier II, France, Janvier 1999

Galinier P, Hertz A (2004) A survey of local search methods for graph coloring. Les cahiers de GERAD G-2004-32, GERAD, Montréal

Hamiez JP, Hao JK (2001) An analysis of solution properties of the graph coloring problem. In: Proc. MIC'2001, 4th metaheuristics international conference, Porto, Portugal, 16–20 July 2001

Hertz A, Jaumard B, de Aragao MP (1994) Local optima topology for the k-coloring problem. Discrete Appl Math 49:257–280

Hordijk W (1995) A measure of landscapes. Technical report 95-045-049, Santa Fe Institute, Santa Fe, New Mexico, USA, May 1995

Johnson DS, Trick MA (eds) (1993) Cliques, coloring, and satisfiability: 2nd DIMACS implementation challenge

Jones T, Forrest S (1995) Fitness distance correlation as a measure of problem difficulty for genetic algorithms. Santa Fe Institute, Working Paper 95-02-022

Kuhn H (1956) Variants of the Hungarian method for assignment problems. Nav Res Logist Q 3:253–258

Merz P, Freisleben B (2000) Fitness landscape analysis and memetic algorithms for the quadratic assignment problem. IEEE Trans Evol Comput 4(4):337–352

Travares J, Pereira FB, Costa E (2008) Multidimentional knapsack problem: A fitness landscape analysis. IEEE Trans Syst Man Cybern, Part B 38(3):604–616

Weinberg B (2004) Analyse et résolution approchée de problèmes d'optimisation combinatoire: application au problème de coloration de graphe. PhD Thesis, Université des Sciences et Technologies de Lille, France